AD_____

Award Number: W81XWH-13-1-0237

TITLE: Impact of Noncoding Satellite Repeats on Pancreatic Cancer Metastasis

PRINCIPAL INVESTIGATOR: David T. Ting, MD

CONTRACTING ORGANIZATION: Massachusetts General Hospital, BOSTON, MA 02114

REPORT DATE: September 2014

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

# REPORT DOCUMENTATION PAGE

*Form Approved*
*OMB No. 0704-0188*



| **1. REPORT DATE** | **2. REPORT TYPE** | **3. DATES COVERED** |
|---|---|---|
| *September 2014* | Annual | 15 Aug 13-14 Aug 14 |

**4. TITLE AND SUBTITLE**
Impact of Noncoding Satellite Repeats on Pancreatic Cancer Metastasis

**5a. CONTRACT NUMBER**

**5b. GRANT NUMBER**
W81XWH-13-1-0237

**5c. PROGRAM ELEMENT NUMBER**

**6. AUTHOR(S)**
David Ting, MD
Daniel A. Haber. M.D. Ph.D. (Mentor)

E-Mail: dting1@partners.org

**5d. PROJECT NUMBER**

**5e. TASK NUMBER**

**5f. WORK UNIT NUMBER**

**7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)**
Massachusetts General Hospital, The

Boston, MA 02114-2621

**8. PERFORMING ORGANIZATION REPORT NUMBER**

**9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)**
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

**10. SPONSOR/MONITOR'S ACRONYM(S)**

**11. SPONSOR/MONITOR'S REPORT NUMBER(S)**

**12. DISTRIBUTION / AVAILABILITY STATEMENT**
Approved for Public Release; Distribution Unlimited

**13. SUPPLEMENTARY NOTES**

**14. ABSTRACT**
*State the purpose, scope, major findings and be an **up-to-date** report of the progress in terms of results and significance*. (Approx. 200 words)

The goal of the project is to understand the role of HSATII satellite repeat expression in pancreatic cancer metastatic potential. An inducible over-expression vector was created and was successfully used in cancer cell lines with evidence of transcriptional changes and increased migratory capability consistent with a role in metastasis. HSATII was also assessed in pancreatic circulating tumor cells (CTCs), which are enriched for metastatic precursors. Initial results find these cells in patients with preneoplastic IPMN lesions suggesting a blood based early detection biomarker. Unexpectedly, a novel reverse transcriptional machinery has been identified with HSATII expression and this results in genomic expansion of these pericentromeric repeats in cancer. This has provided a new understanding of HSATII regulation and function, which has adjusted the goals of the project to address these new findings.

**15. SUBJECT TERMS :** nothing listed

| **16. SECURITY CLASSIFICATION OF:** | | | **17. LIMITATION OF ABSTRACT** | **18. NUMBER OF PAGES** | **19a. NAME OF RESPONSIBLE PERSON** |
|---|---|---|---|---|---|
| **a. REPORT** | **b. ABSTRACT** | **c. THIS PAGE** | | | USAMRMC |
| U | U | U | UU | *49* | **19b. TELEPHONE NUMBER** *(include area code)* |

## Table of Contents

## Introduction

Pancreatic cancer remains one of the most deadly cancers where the vast majority of patients are diagnosed too late and conventional therapies have largely been ineffective, making early detection and novel drug targets greatly needed. Recent studies have shown the expression of a significant portion of genomic regions previously thought to be transcriptionally silent. Satellites are regions of the genome that are highly repetitive and normally their expression is suppressed by heterochromatin, however, their expression was found to be abundant in a wide variety of cancers. The goal of this research is to understand the cellular and molecular impact of satellite RNA in cancer cells and to test the utility of these highly specific and abundant transcripts as novel biomarkers for early detection.


## Keywords

cancer genetics, satellite repeats, metastasis, circulating tumor cell, pancreatic cancer


## Overall Project Summary

All tools and were developed for each aim and the initial experiments designed to understand the impact of satellite expression in cancer cell lines demonstrated increased migratory function in cell lines with inducible HSATII overexpression (See Aim 1 below). However, in parallel experiments, we sought to understand the endogenous expression of HSATII in cancer cells. We had noted in prior work in both mouse and human cancer cell lines that there is suppression of satellite repeat expression in standard in vitro adherent culture. Massive satellite expression could be induced upon inoculation of tumor cells in immunodeficient mice suggesting in vivo environmental stimuli provided critical signals for the expression of these repeats. We performed a number of perturbations in vitro to induce HSATII expression including hypoxia, UV radiation, demethylation, starvation, and growth in non-adherent conditions. Only growth in non-adherent conditions (as tumor spheres or in soft agar) was sufficient to induce HSATII expression in multiple cancer cell lines including pancreatic and colon cancers (see attached submitted manuscript). In addition, we unexpectedly found that a significant portion of the HSATII transcripts were in fact complementary DNA pointing towards a reverse transcriptional machinery, which has not been described for human satellite repeats. Because of this unexpected finding, we focused our efforts over the last reporting period in validating this novel finding and to understand the significance of this phenomenon in cancer function. Since there was growing evidence of LINE1 retrotransposition activity in colon cancer [1], we used colon cancer as a model to best study this novel reverse transcriptional mechanism. Through a number of biochemical experiments we believe reverse transcriptional machinery is highly active and specific for satellite repeats in human cells. These RNA derived DNAs (rdDNA) are found in primary tumors, xenografts, and tumorspheres in large amounts and appear to be used as templates for elongating the pericentromeric regions from where the originate from. We validated the DNA expansion of HSATII regions in our xenograft models and find 50% of primary colon cancers with significant copy number gains of HSATII as determined by whole genome sequencing. Though we do not yet know the functional consequences of HSATII expansions in the tumor genome, there is other work that suggests that this may stabilize centromeric regions to tolerate increased replication stress common in cancer cells. For full details, see attached manuscript that has been submitted and is in peer review.

**Aim 1: Evaluation of Satellite expression on transcriptional profiles**

Task 1. Development of satellite expressing cell lines: A doxycycline inducible HSATII vector was created containing a segment of HSATII that was approximately 800 bp in length. This vector was successfully transduced into human cancer cell line SW620. Induction of HSATII expression was clearly evident by the addition of doxycycline and confirmed by RNA-ISH (Fig. 1) and northern blot.

Task 2. Effects of satellite on expression patterns: Satellite induction was performed with doxycycline and cell lines were evaluated for expression pattern changes using the Helicos single molecule sequencing platform. A total of 126 genes were differentially expressed in HSATII induced cell lines compared to GFP induced cell line controls. Gene ontology of these genes did not identify any major signaling pathways or other known expression profile enrichment. Notably, 46% of these genes are upregulated in human brain tissue consistent with our prior correlation of satellite expression to neural genes.

*Although forced over-expression of HSATII did result in transcriptional and functional changes, the reverse transcriptional effect as described above was not seen in cells. Given the inconsistency of HSATII transcript behavior by endogenous loci and our lentiviral construct led us to reconsider the relevance of studying HSATII through a conventional overexpression lentiviral based system.*

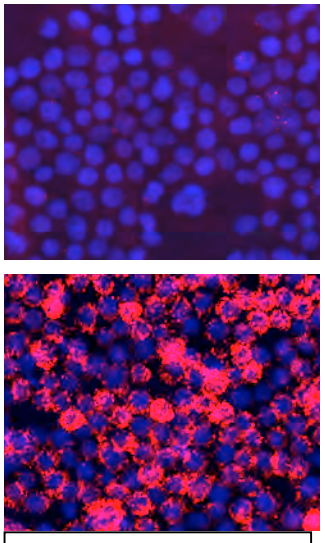


Fig. 1: SW620 HSATII inducible cell line without (top) and with doxycycline (bottom). HSATII RNA-ISH (red) and DAPI nuclear stain (blue)

Task 3. Effects of satellites on epigenetic marks: This task was not pursued due to changes in our aims as noted above.

**Aim 2: Evaluation of Satellite expression on metastatic potential**

Task 1. Effects on adherent culture: Effects of HSATII induction in HSATII cell lines was performed showing some changes in morphology of the cell line and increased migratory function. However, there were no appreciable effects on cell growth and a potentially negative effect on tumor sphere formation with HSATII overexpression.

Task 2. Effects on xenograft tumors: Due to changes in priorities as noted above, we have deferred xenograft tumor formation with HSATII inducible cell lines. Instead, we are focusing on the mechanism of massive HSATII expression in tumorigenesis.

Task 3. Effects on CTCs: Due to changes in priorities as noted above, we have deferred CTC analysis. However, based on HSATII expression in cell lines we note that there is increased migratory function, which may increase CTC formation. However, the reduction of tumor sphere capability suggests that HSATII induction may also reduce CTC viability in circulation due to anoikis. We have decided to focus on understand the endogenous induction of HSATII and its relationship to tumorigenesis and metastasis given the inducible HSATII expression does not create the same magnitude or quality of HSATII expression seen in the endogenous cell line setting. Recent, analysis of mouse pancreatic CTC data has identified elevated repeat expression in CTCs compared to matched primary tumors pointing towards a correlation of increased satellite expression in CTCs compared to primary tumor cells. Human HSATII expression in CTCs (See Aim 3) has also shown increased detection sensitivity of CTCs again pointing towards a relationship of satellite expression to the metastatic process.

**Aim 3: Evaluating Satellites as a novel CTC Biomarker**

Task 1. Optimization of RNA-ISH assay for HSATII in CTCs: HRPO approval and initiation of testing HSATII ISH in clinical samples has been done over the last year. Initial testing of RNA-ISH on the 3$^{rd}$ generation IFD CTC-chip has been completed. Optimization of automated imaging analysis is still ongoing, but the assay appears to be working well on the new CTC device.

Task 2. Comparative analysis of HSATII RNA-ISH versus CK/EpCAM Immunofluorescence CTC enumeration assays: We have decided to test the assay on a high risk patient population who are being monitored for preneoplastic cystic lesions of the pancreas known as intraductal papillary mucinous neoplasm (IPMN). Approximately 20-30% of these patients will develop pancreatic cancer and therefore, an early detection blood based biomarker is of high clinical importance. We have done an initial test of 5 patients with IPMN seen at the MGH and compared CK immunofluorescence (IF) to HSATII ISH in a split blood sample run on the same IFD CTC-chip. Notably, HSATII ISH is detecting far more events (Table 1) than CK IF, which provides the sensitivity and dynamic range needed for a prognostic biomarker. These results are encouraging and we are monitoring these patients for clinical outcomes to see if high HSATII ISH CTCs are at increased risk from the development of PDAC. In parallel, we are testing a cohort of patient with resectable pancreatic cancer to again determine if HSATII ISH CTCs can increase the sensitivity of the assay. These two patient populations we are testing would provide the foundation for using this assay as an early detection strategy in pancreatic cancer.

Table 1: Candidate CTC counts from IPMN patients

| Patient ID | CK IF | HSATII ISH |
|---|---|---|
| IPMN 1 | 16 | 86 |
| IPMN 2 | 0 | 75 |
| IPMN 3 | 0 | 6 |
| IPMN 4 | 0 | 21 |
| IPMN 5 | Failed | 240 |

## Key Research Accomplishments

1) Developed HSATII inducible cell line demonstrating some effects on migration caused by HSATII overexpression highlighting a functional effect of this non-coding RNA in cell lines.

2) Identification of loss of adherence as the sole environmental perturbation that induces HSATII expression across multiple human cancer cell lines, which has provided new insight into the regulation of HSATII in cancer

3) Discovery of a novel reverse transcriptional mechanism that has never been described for satellite repeats in humans. This process can be inhibited by small molecule NRTIs, which offers a potential new cancer therapy.

4) Identification of HSATII genomic expansion as a common feature across epithelial cancers, the functional significance of which remains to be determined.

5) Preliminary evaluation of HSATII as a CTC biomarker in patients at high risk of developing pancreatic cancer indicates much improved sensitivity of detection, which may provide a blood based early detection modality.

## Conclusion

The massive expression of satellite repeats in virtually all epithelial cancers was an unexpected finding with implications as a cancer diagnostic and also as a new unappreciated phenomenon in cancer biology. Our original plans to study overexpression of HSATII in cancer cells demonstrated an increase in migratory function suggesting a link with the metastatic cascade, which is the main cause of mortality in solid malignancies. Increased levels of satellites in circulating tumor cells (CTCs) supports a relationship of HSATII with metastasis and more importantly this may prove to be an blood based early detection diagnostic for pancreatic cancer. In our pursuit, to understand the biological regulation and function of satellites, we have discovered a novel reverse transcriptional mechanism that expands satellite repeat regions in the genome, which may have implications in tumor cell survival. In summary, we have made significant progress in understanding the mechanistic underpinnings of these repetitive elements in cancer. This has adjusted our experimental goals of the project and we are planning to focus on two major questions based on our new findings.

The modified aims are as follows for this project:

**Aim 1: What is the mechanism of endogenous HSATII expression in cancer cells when grown in non-adherent conditions?** We will use a combination of RNA-sequencing to evaluate for transcriptional signatures that can highlight the pathways that are changed in cancer cells grown in 2D or 3D environments. We will evaluate for conjugate changes of protein expression with appropriate western blot analyses.

**Aim 2: What are the effects of perturbing HSATII reverse transcription with anti-sense nucleic acids or reverse transcriptase inhibitors?** We have already identified the nucleoside reverse transcriptase inhibitor ddC as a small molecule inhibitor of HSATII reverse transcription. Initial data indicates there are anti-proliferative effects of ddC in cancer cell lines. We will evaluate ddC and anti-sense locked nucleic acids as methods for inhibiting this process and evaluate their effects in cancer cells *in vitro* and *in vivo*. This has implications as a novel therapeutic target that can help provide new avenues to treat cancer.

## Publications, Abstracts, and Presentations

**Publications (related)**

1. Yu M, Bardia A, Aceto N, Bersani F, Madden MW, Donaldson MC, Desai R, Zhu H, Comaills V, Zheng Z, Wittner BS, Stojanov P, Brachtel E, Sgroi D, Kapur R, Shioda T, **Ting DT**, Ramaswamy S, Getz G, Iafrate AJ, Benes C, Toner M, Maheswaran S, and Haber DA, Cancer therapy. Ex vivo culture of circulating breast tumor cells for individualized testing of drug susceptibility. *Science*, (2014); 345(6193): 216-20.

2. **Ting DT** and Ryan DP, The wide gulf between stage III and stage IV colon cancer. *Lancet Oncol*, (2014); 15(8): 785-6.

3. Luo X, Mitra D, Sullivan RJ, Wittner BS, Kimura AM, Pan S, Hoang MP, Brannigan BW, Lawrence DP, Flaherty KT, Sequist LV, McMahon M, Bosenberg MW, Stott SL, **Ting DT**, Ramaswamy S, Toner M, Fisher DE, Maheswaran S, Haber DA. Isolation and molecular characterization of circulating melanoma cells. *Cell Reports* (2014); 7(3):645-53.

4. Javaid S, Zhang J, Anderssen E, Black JC, Wittner BS, Tajima K, **Ting DT**, Smolen GA, Zubrowski M, Desai R, Maheswaran S, Ramaswamy S, Whetstine JR, Haber DA. Dynamic chromatin modification sustains epithelial-mesenchymal transition following inducible expression of Snail-1. *Cell Reports* (2014); 5(6):1679-89.

**Publications Submitted**

Bersani F, Lee E, Kharchenko PV, Xu AW, Xega K, Brannigan BW, Wittner BS, Ramaswamy S, Park PJ, Maheswaran S, **Ting DT***, Haber DA*. "Pericentromeric satellite repeat expansions through RNA-derived DNA intermediates in cancer". * Denotes co-corresponding authors

**Oral Presentations**

**Ting DT**. Diversity of Circulating Tumor Cells in a Mouse Pancreatic Cancer Model Identified by Single Cell RNA Sequencing. AACR 2014. San Diego, CA. April 2014

## Inventions, Patents and Licenses

Nothing to report

## Reportable Outcomes

Nothing to report

## Other Achievements

Nothing to report

## References


1. Lee E, Iskow R, Yang L, Gokcumen O, Haseley P, Luquette LJ, 3rd, Lohr JG, Harris CC, Ding L, Wilson RK, Wheeler DA, Gibbs RA, Kucherlapati R, Lee C, Kharchenko PV, and Park PJ, Landscape of somatic retrotransposition in human cancers. *Science*, (2012); 337(6097): 967-71.


See additional references in attached submitted manuscript

## Appendices

See attached manuscript that has been submitted and is under review

**Pericentromeric satellite repeat expansions through RNA-derived DNA intermediates in cancer**

Francesca Bersani,[1] Eunjung Lee,[4,5] Peter V. Kharchenko,[4,6] Andrew Wei Xu,[4] Mingzhu Liu,[1,8] Kristina Xega,[1] Brian W. Brannigan,[1] Ben S. Wittner,[1] Sridhar Ramaswamy,[1,2] Peter J. Park,[4,5,7] Shyamala Maheswaran,[1,3] David T. Ting,[1,2*] Daniel A. Haber[1,2,8*]

[1] Massachusetts General Hospital Cancer Center and Departments of [2] Medicine and [3] Surgery, Harvard Medical School, Charlestown, MA 02129.

[4] Center for Biomedical Informatics, Harvard Medical School, Boston, MA 02115.

[5] Division of Genetics, Brigham and Women's Hospital, Boston, MA 02115.

[6] Hematology/Oncology Program, Children's Hospital, Boston, MA 02115.

[7] Informatics Program, Children's Hospital, Boston, MA 02115.

[8] Howard Hughes Medical Institute, Chevy Chase, MD 20815.

[*] To whom correspondence should be addressed:

D.A.H (haber@helix.mgh.harvard.edu) and D.T.T. (dting1@mgh.harvard.edu)

**Comprehensive transcriptome analysis has revealed that close to three quarters of the human genome is pervasively transcribed, whereas less than 2% is ultimately translated into proteins[1]. Among the recently appreciated non-coding RNAs (ncRNAs) in eukaryotes are classes of pericentromeric satellite repeats[2]. While satellite repeats are characterized by extreme interspecies sequence diversity, they share a conserved function as core centromere-building elements, thereby stabilizing interactions with DNA-binding proteins, sustaining kinetochore formation, and driving chromosomal segregation during mitosis[3]. Like other heterochromatic repetitive elements once considered transcriptionally inactive, fine modulation of transcription in these regions has recently been shown to be essential to maintaining heterochromatin architecture and chromosomal integrity[4]. Aberrant massively elevated transcription of the HSATII satellite repeat has been found in a wide variety of epithelial tumours[5]. In deriving experimental systems to study its deregulation, we observed that HSATII expression is induced in colon cancer cells cultured as xenografts or under non-adherent conditions *in vitro*, but is rapidly lost in standard 2D cultures. Remarkably, induction of endogenous HSATII RNA under these physiological conditions, as well as introduction of synthetic HSATII transcripts, generate complementary DNA intermediates in the form of DNA:RNA hybrids, through a reverse transcriptase mediated process. Single molecule sequencing of mouse tumour xenografts and primary human tumours shows that HSATII RNA-derived DNA (rdDNA) molecules are stably reincorporated within pericentromeric loci, leading to progressive expansion of these regions. A broad analysis of whole genome sequencing data reveals that HSATII copy number gain is a common feature in primary colon tumours. Our observations suggest that a RNA-driven HSATII genomic expansion**

**influences the architecture of pericentromeres in cancer, pointing to a previously unappreciated level of genetic instability in human malignancies.**

The highly repetitive nature of satellites precludes their precise quantitation and qualitative analysis using PCR-based RNA sequencing approaches. We previously showed that PCR-independent single molecule next-generation sequencing (digital gene expression profiling, DGE; Helicos/SeqLL) is uniquely sensitive and quantitative in measuring these repeats[6], although it is not suited to high throughput analysis. To enable experimental models for the study of HSATII deregulation, we first designed a modified Northern blot HSATII assay (Fig. 1a, b). HSATII satellite transcripts encompass arrays of variable lengths derived from multiple different genomic locations[7], thus Northern blotting generated a pattern of bands ranging from ~30 nt to greater than 800 nt in size (Fig. 1b), consistent with that reported for other repeats[8-10]. Quantitation of DGE profiles and Northern blot signal intensity for matched primary gastrointestinal cancer specimens were highly correlated (Fig. 1b, c).

We observed that human colorectal cancer cell lines do not express HSATII under standard *in vitro* adherent (2D) culture conditions, but strongly upregulate its expression when grown as tumour xenografts (Fig. 1d). To define specific experimental conditions that modulate HSATII expression within tumours, we tested multiple stimuli associated with cellular stress and tumourigenesis (Extended Data Fig. 1a-d). Remarkably, only culture under non-adherent conditions, as 3D tumour spheres in solution or in soft agar, led to robust induction of HSATII in five colorectal cancer cell lines, as detected by Northern blotting (Fig. 1e and Extended Data Fig. 1c). This phenomenon was also evident using RNA-in situ hybridization (ISH), an imaging assay that does not involve any nucleic acid extraction (Fig. 1f), thus linking HSATII expression to anchorage-independent growth. A sixth colorectal cancer line, COLO205, noteworthy

for its growth as a mixed population of adherent and floating cells, was unique in expressing HSATII RNA at baseline under standard culture conditions (Fig. 1e). Notably, the elevated RNA levels detected in tumour spheres and xenografts were rapidly lost upon replating cells in adherent 2D culture (Extended Data Fig. 1e). HSATII transcripts were present in both sense and antisense orientation and primarily localized to the nuclear compartment (Extended Data Fig. 1f, g).

Unexpectedly, we observed that a fraction of the xenograft-induced HSATII sequences present within medium/small molecular weight nucleic acids (Trizol extraction) was sensitive to DNase I (Fig. 2a). We excluded the possibility that this signal resulted from contamination with cellular genomic DNA (gDNA) by performing a detailed series of controls (see Supplementary Data 1 and Extended Data Fig. 2). Thus in cells that overexpress the HSATII satellite repeat, the deregulated HSATII transcripts coexist with matched DNase I-sensitive fragments, possibly deriving from DNA:RNA hybrids. This hypothesis would be of considerable interest, since such hybrids may be originated from reverse transcription (RT) of satellite transcripts, a phenomenon known to occur with better characterized repetitive elements, such as LINEs, SINEs and LTR retrotransposons[11], whose activity in cancer cells has been reported to increase genomic instability[12]. To test the possibility of HSATII RNA-to-DNA conversion, we first developed an assay to directly introduce synthetically produced HSATII RNA generated by *in vitro* transcription (IVT) into 2D-cultured 293T cells that lack endogenous HSATII expression. To generate an appropriate IVT transcript, we used a pericentromeric HSATII locus at chromosome 10q11 (HSATII-chr10), whose expression is highly induced in cancer[5]. Since IVT using T7 polymerase relies on a PCR-generated DNA template as starting material, we included multiple controls to ensure the absence of DNA template contamination within the single strand IVT product itself, as well as any genomic HSATII

sequences in the cellular extracts (see details in Supplementary Data 2 and Extended Data Fig. 3).

To assess the formation of a DNA:RNA hybrid in cells transfected with single stranded IVT HSATII RNA, we subjected nucleic acid extracts to treatment with RNase H, which specifically digests the RNA moiety of DNA:RNA hybrids, but does not affect either single stranded RNA or the DNA component of DNA:RNA hybrids (Fig. 2b). Indeed, RNase H treatment caused a strong reduction in the Northern blot signal identified for the HSATII sense sequence (HSATII S; Fig. 2c), indicating that part of the IVT-produced RNA is complexed with a complementary DNA strand, consistent with the generation of a RNA-derived-DNA (rdDNA). Transfection of comparable amounts of IVT GFP RNA produced the expected RNA signal but showed no significant sensitivity to RNase H (Fig. 2d), nor did introduction of GFP RNA lead to the induction of HSATII satellite RNA (Fig. 2c). Together, these results indicate that ectopically introduced single stranded HSATII RNA is unique in generating complementary DNA within transfected cells.

To further validate these results, we made use of the S9.6 monoclonal antibody which is highly specific in its recognition of DNA:RNA hybrids[13-16]. We established a DNA:RNA hybrid immunoprecipitation (DRIP) assay using real-time PCR (HSATII-chr10 qPCR) of S9.6 immunoprecipitates, which was applied to nucleic acids from untransfected or IVT HSATII RNA-transfected cells. Samples were subjected to complete DNAse I digestion (which removes all double strand DNA but does not affect DNA:RNA hybrids), followed by treatment with RNase H or control, and finally DRIP analysis (Fig. 2e). HSATII DNA:RNA duplexes were present only in 293T cells transfected with HSATII RNA. RNase H treatment effectively abolished the immunoprecipitation of HSATII DNA:RNA hybrids (Fig. 2f).

We then assessed the presence of endogenous HSATII DNA;RNA hybrids by applying the DRIP assay to COLO205 cancer cells, whose anchorage-independent growth is associated with baseline expression of HSATII transcripts (Fig. 1e). Indeed, RNase H-sensitive DNA:RNA HSATII hybrids were immunoprecipiated by DRIP assay in COLO205 cells grown under baseline conditions (Fig. 2g). To evaluate the consequences of a RT inhibitor on formation of these hybrids, we tested the effect of the nucleoside analog RT inhibitor (NRTI) 2',3'-dideoxycytidine (ddC) in COLO205 cells (Fig. 2e). Notably, ddC is very poorly incorporated by replicative polymerases[17], while displaying high specificity for different classes of RT[18], and it is well suited to the HSATII template which is highly CT-rich in sequence. Indeed, ddC treatment significantly reduced the levels of endogenous HSATII DNA:RNA hybrids in COLO205 cells, as measured by the DRIP assay (Fig. 2g). A decrease in total HSATII rdDNA following treatment of COLO205 cells with ddC was also evident using HSATII-chr10 qPCR analysis of RNase A-treated Trizol extracts (Extended Data Fig. 3i). Together, these observations are consistent with RT activity in HSATII-expressing cells, contributing to the generation of DNA:RNA structures derived from the satellite transcripts.

HSATII-derived rdDNA fragments present in the nuclear fraction (Extended Data Fig. 1g) may give rise to extrachromosomal elements, or may be integrated at chromosomal loci, leading to stable expansion of HSATII genomic sequences. By analogy, RT of LINE-1 transcripts, followed by their retrotransposition and reintegration at chromosomal loci has been described in epithelial cancers, including colon carcinoma[19]. To address this possibility, we first analysed the dynamics of HSATII RNA and DNA level changes using single molecule sequencing (Helicos/SeqLL) in SW620 colon cancer cells transitioned from 2D *in vitro* culture conditions to growth as mouse xenografts and vice versa. As expected, the number of HSATII RNA reads was minimal when cells were cultured in 2D conditions, induced 360-fold as the cells gave rise to

xenografts in mice, and then promptly downregulated as xenograft-derived tumour cells were returned to *in vitro* 2D cultures (Fig. 3a). Remarkably however, total cellular HSATII DNA copy number, which was already abundant at baseline, increased as much as 25-fold as 2D-cultured cells were transitioned to xenografts and then remained stably amplified despite return of cells to 2D culture *in vitro*. As a control, we analysed the GSATII satellite which is structurally similar to HSATII but whose expression is not deregulated in cancer[5]. SW620 cells showed negligible GSATII changes, either in RNA or DNA content, as cells transitioned between 2D *in vitro* and xenograft culture conditions (Fig. 3a and Extended Data Table 1). DNA fluorescent in situ hybridization (FISH) applied to cells with amplified HSATII DNA sequences did not reveal detectable extrachromosomal elements, nor was any hybridization signal visible outside the five chromosomal loci known to harbour long arrays of pericentromeric HSATII (Fig. 3b). Consistent with the FISH data, analysis of HSATII gDNA reads from the SW620 xenografts, using single molecule sequencing, showed that the additional HSATII sequences were distributed among the various endogenous pre-existing HSATII pericentromeric loci (Fig. 3c and Extended Data Fig. 4).

To model HSATII DNA copy gain over time and as a function of tumour progression, we serially transplanted SW620 cells as xenografts over multiple generations of mice. Progressive amplification of HSATII gDNA was evident over ten successive rounds of *in vivo* tumour initiation. This was assessed using a real-time PCR-based copy number variation (CNV) assay at the two highest density HSATII pericentromeric regions on chromosome 16q (HSATII-chr16-1 and -2; Fig. 3d, e). An adjacent chromosomal region showed no xenograft transplantation-associated copy number changes, ruling out nonspecific gains in the 16q chromosomal arm or in ploidy (Fig. 3f). Rather than showing a discrete locus duplication event at a specific time in tumour progression, the pericentromeric genomic loci appear to demonstrate a gradual

increase in HSATII gene copy number over time, all within preexisting satellite domains. Such a timeline would be consistent with the possibility of multiple rdDNA-mediated reintegration events.

To determine whether HSATII copy number gains occur in primary human colon cancer, we analysed CNV in 10 pairs of tumour and matched adjacent normal tissue, focusing on the chromosome 16q (HSATII-chr16-1 and -2) loci. After correcting for chr16q arm loss or gain, significantly increased HSATII copy number was evident at either or both of the two independent HSATII loci tested in 5/10 (50%) colon cancers (Fig. 4a-c). Among other cancers analysed, similar HSATII gene copy gain was evident in 5/13 (38%) kidney cancers (Extended Data Fig. 5a).

To date, the complexity of sequencing and assembling genomic reads from highly repetitive sequences has precluded their detailed analysis in cancer through genome wide databases[20]. To extend our study of focal HSATII loci in selected tumour types to a genome-wide survey of all such satellite repeats, we used a satellite CNV algorithm to undertake computational analyses of whole genome sequencing (WGS) from both TCGA and a published whole genome sequence database[21]. In fully annotated genomic sequences of 38 colorectal cancers, 20 (53%) had statistically significant genomic gain of HSATII compared with their matched normal germline (Fig. 4d, e). HSATII amplification in these tumours was only partly overlapping with the changes affecting other satellite repeats (ALR/Alpha, CATTCn, TAR1, GSATII and BSR/Beta). We note that the read depth obtained on these samples (>30X) provides high confidence CNV calls, but the intrinsic difficulty in PCR-based sequencing and subsequent alignment of repetitive sequences still poses considerable challenges. The high-stringency algorithm that we applied is therefore likely to underestimate the true magnitude of gene copy changes. Correction of these data for large genomic alterations, comparable in size with HSATII stretches, returned consistent results (Extended Data

Fig. 5b). Taken all together, our data show that gene copy gains at HSATII-encoding pericentromeric repeats are a common feature of colorectal cancers.

In summary, we have shown that the normally silenced HSATII class of pericentromeric satellites is specifically expressed in colon cancer cells cultured under non-adherent conditions *in vitro* or grown as tumour xenografts *in vivo,* and that this phenomenon is associated with the production of RT-dependent DNA:RNA hybrids and expansion of HSATII pericentromeric chromosomal domains. These observations suggest a retro-insertion mechanism for pericentromeric satellites that has previously been associated with LINEs and other reverse transcriptase-encoding repeats and it points to an unexpected degree of cancer-associated genomic instability at pericentromeric repeats, which constitute critical structural components of eukaryotic chromosomes (Extended Data Fig. 6). While the RT-directed expansion of centromeric sequences has not been described in human cells, there is ample precedent for retro-element-mediated integration of centromeric repeats in other species[22,23]. In mammalian cells, reverse transcription of other major repetitive elements is well established, but RT-mediated integration of DNA sequences within centromeres has not been reported, with the exception of the marsupial tammar wallaby, whose exceptionally short centromeres harbour signatures of retroviral insertions alongside domains of satellite-rich sequences[24]. HSATII reverse transcription may be mediated by a variety of endogenous RTs encoded by retroviral elements, LINE-1, telomerase, or a yet to be identified enzyme. While we favour RT-mediated chromosomal reintegration as the mechanism for pericentromeric expansion, other possible mechanisms for HSATII copy number gain include unequal chromosomal recombination, rolling circle amplification, gene conversion, replication slippage, and epigenetic modifier-induced site specific copy gain. Further studies to identify the specific pathways mediating HSATII genomic amplification

in cancer cells may provide insight into the instability and regulation of these critical structural domains.

At present, we can only speculate on the consequences of pericentromeric expansion during cancer progression. Physical stretching of the kinetochore predisposes centromeric regions to DNA breaks[25] and the double-strand breaks associated with replication stress are known to occur at high frequency within satellite repeat-enriched genomic regions[26]. Pericentromeric heterochromatin is also essential for initiation of the mitotic spindle, whose disruption is associated with chromosomal defects in human tumours[27]. Thus, pericentromeric expansion may contribute both to intra-chromosomal stability as well as to integrity of the mitotic spindle and chromosomal segregation. While synthetic reconstitution models will be required to test these hypotheses, we speculate that HSATII expression followed by amplification of these genomic loci may result in attenuation of chromosomal instability associated with cancer progression. Finally, while much remains to be learned about the mechanisms and consequences of pericentromeric deregulation in tumours, the very high level expression of these satellite transcripts in virtually all human epithelial cancers, combined with their appropriation of a reverse transcriptase-dependent pathway for pericentromeric expansion may point to novel therapeutic opportunities.

## ACKNOWLEDGEMENTS

We are grateful to M. Miri from the MGH Tissue Repository for providing pathological specimens. We also thank N.J. Dyson and M.R. Motamedi for critically revising the manuscript as well as R. Taulli and all lab members for helpful discussions. This work was supported by Howard Hughes Medical Institute (D.A.H), National Foundation for Cancer Research (D.A.H.), NCI R01CA129933 (D.A.H., F.B.), the Burroughs Wellcome Trust (D.T.T.), NIH K12CA087723-11A1 (D.T.T.), Department of Defense W81XWH-13-1-0237 (D.T.T.), the Warshaw Institute for Pancreatic Cancer Research (D.T.T.), the Verville Family Pancreatic Cancer Research Fund (D.T.T.), Susan G. Komen for the Cure KG09042 (S.M.) and NCI Federal Share Program and Income (S.M.).

## AUTHOR CONTRIBUTION

F.B., D.T.T., S.M. and D.A.H. conceived and designed the study; F.B., K.X. and B.B. conducted experiments; E.L., P.V.K, A.W.X. and P.J.P. developed analytical tools and performed computational analysis; B.S.W. and S.R. provided bioinformatic support; M.L. supplied technical expertise; F.B. and D.T.T. analysed and interpreted data; F.B., D.T.T., S.M. and D.A.H. wrote the manuscript. All authors reviewed and revised the manuscript.

## COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

## METHODS

### Cell lines and cellular assays

All colon cancer cell lines (ATCC) were cultured in RPMI-1640 + 10% FBS + 1% Pen/Strep (Gibco/Life Technologies). DMEM (Gibco/Life Technologies) with the same supplements was used for 293T cells (ATCC). All cells were tested for mycoplasma contamination giving negative results. For drug treatment, cells were cultured in the presence of either DMSO or 100 μM 2',3'-dideoxycytidine (ddC, Sigma) for two weeks.

For tumour sphere assays, 1 x $10^3$ cells were plated as single cell suspension in ultralow attachment 6-well plates (Corning) and grown in serum-free RPMI medium supplemented with 20 μl/ml B27 (Invitrogen/Life Technologies), 20 ng/ml EGF (Invitrogen/Life Technologies) and 20 ng/ml bFGF (Invitrogen/Life Technologies). Fresh medium was added every 3 days. Spheres were collected at day 10.

For soft agar anchorage-independent growth assays, cells were suspended in 0.4% low-melting agarose (Sigma Aldrich) in 10% RPMI at a density of 2 × $10^4$/well, plated on a layer of 0.8% agarose in 10% RPMI in 6-well plates and cultured for 2 weeks.

### Xenografts studies and human tissues

Mouse xenograft studies were performed according to an animal protocol approved by the MGH Subcommittee on Research Animal Care. Cells (1 x $10^6$) were injected in the flank of six-week old female nude mice (Charles River Laboratories). Tumours were collected 3 weeks later. Cell cultures were reestablished from xenografts after collagenase IV (Sigma) digestion in 1X HBSS (Gibco/Life Technologies) and filtration through a 70 μm cell strainer. For serial transplantation, tumours were minced and a small piece (~2 mm x 2 mm) was transferred subcutaneously into a recipient mouse.

Human normal and tumour tissues were obtained from the Massachusetts General Hospital according to an IRB-approved protocol and following informed consent from all patients. Total RNA from normal human pancreas was purchased from Clontech.

**RNA/DNA extraction and nuclease treatment**

Total RNA was extracted with Trizol (Ambion/Life Technologies) according to standard procedures. Where indicated and in all IVT experiments, Phase Lock Gel tubes (5 Prime) were used for Trizol-based extraction. Nuclear/cytoplasmic RNA extracts were obtained using Paris Kit (Ambion/Life Technologies) pursuant to the protocol guidelines. DNA was extracted with DNeasy Blood & Tissue Kit (Qiagen) according to instructions including the optional RNase A digestion step. Total nucleic acid (TNA) extraction was done using MasterPure Complete DNA and RNA Purification Kit (Epicentre) according to manufacturer's protocol and followed by complete removal of contaminating DNA by DNase I digestion.

Nuclease treatments (0.5 U/µl DNase I, Roche; 0.5 U/µl RNase H, New England BioLabs; 10 ng/µl RNase A, Ambion/Life Technologies) were carried out following manufacturer's guidelines.

**Northern blot**

For Northern blot analysis, total RNA (5 µg) before or after nuclease treatment was electrophoresed in a 4% or 8% polyacrylamide-urea gel and transferred by electroblotting onto Hybond-N+ membrane (Amersham/GE Healthcare). Hybridization was performed with the following $^{32}$P-labeled DNA oligos: *anti-HSATII S*, 5'-cattcgattccattcgatgat-3'; *anti-HSATII AS* (also indicated in the text and figures as *HSATII*), 5'-atcatcgaatggaatcgaatg-3'; *anti-GFP S*, 5'-cgcgcttctcgttggggtctt-3'; *anti-U6*,

5’-ttgcgtgtcatccttgcgcagg-3’. Relative signal density was quantified with Image J software.

### *In vitro* transcription and transfection

HSATII was amplified from the genome with the following primers: FOR, 5’-cattccattccattagatgattcca-3’; REV, 5’-tgaatggaataatccttgaacggaa-3’ and inserted into pCRII-TOPO vector (Invitrogen/Life Technologies) *via* TOPO-TA Cloning (Invitrogen/Life Technologies). GFP was subcloned into the BamHI-SalI sites of pSuper vector (Oligoengine). Both fragments were amplified by PCR with M13 for/rev primers and subjected to *in vitro* transcription with T7 RNA polymerase following the T7 transcription kit recommendations (Roche). This included a Turbo DNase digestion step after RNA transcription to prevent DNA carryover. RNA (12.5 µg) was then transfected using Lipofectamine 2000 (Invitrogen/Life Technologies) into a 6 cm tissue culture dish containing sub-confluent 293T cells according to manufacturer’s instructions. Cells were collected 24 h post-transfection for nucleic acid extraction.

### DNA:RNA hybrid immunoprecipitation (DRIP)

DRIP was performed with minor modifications of a previously published protocol[16]. Briefly, total nucleic acid (TNA) extraction followed by complete DNAse I and RNase H (where indicated) digestion was done as described above. All incubations and wash steps were carried out in 10 mM EDTA/PBS buffer supplemented with RNaseOUT (Invitrogen/Life Technologies). Anti-DNA:RNA Hybrid [S9.6] antibody (ENH001, Kerafast; 12.5 µg/sample) was pre-adsorbed on Protein G Dynabeads (Invitrogen/Life Technology; 30 µl/sample) for 2 h at 4°C. After washing, Protein G-bound antibody was incubated overnight at 4°C with 10 µg of sample, and 1% input was set aside as control. Beads were then washed five times for 10 minutes and immunocomplexes were eluted

in 200 µl of 10 mM EDTA/0.6% SDS followed by standard ethanol precipitation. Real-time PCR was performed with the *HSATII-chr10 (IVT)* primer pair for HSATII IVT-transfected 293T cells and the *HSATII-chr10* primer pair for COLO205 cells (see details below). DRIP fraction Ct values were normalized against input fractions (1%) and fold change was calculated for each sample as the $2^{(-\Delta\Delta Ct)}$ of the normalized "RNase H-" DRIP Ct values over the normalized background ("RNase H+" DRIP fraction), which was set at 1. All experiments were performed independently three or more times with consistent results.

**qPCR**

All real-time PCR (qPCR) assays were done using Power SYBR-Green PCR Maser Mix on an ABI/PRISM 7500 platform (all reagents were from Applied Biosystems/Life Technologies). The following primers were employed: *HSATII-chr10 (IVT) for*, 5'-caccaaacggaaaaaaacgga-3'; *HSATII-chr10 (IVT) rev*, 5'-gagctcggatccactagtaacg-3'; *HSATII-chr10 for*, 5'-gcattcaattcattagatgacgg-3'; *HSATII-chr10 rev*, 5'-ccttgacccgaatgcaatca-3'. In the RNase A digestion experiment followed by HSATII-chr10 qPCR, Ct values were normalized against the matched untreated samples before applying the $2^{(-\Delta\Delta Ct)}$ analysis method.

For copy number variation assessment on purified DNA, the following primer pairs were used: *HSATII-chr16-1 for*, 5'-ccattcgttaatgcctttcg-3'; *HSATII-chr16-1 rev*, 5'-cacgaatggaatcattgtcg-3'; *HSATII-chr16-2 for*, 5'-tccattcgaggattccactc-3'; *HSATII-chr16-2 rev*, 5'-caaagggaagcaaaggaaatc-3'; *16q for*, 5'-ggggtaaaagggcatgtttt-3'; *16q rev*, 5'-ctgaagaagcccactcaagg-3'; *β-Actin for*, 5'-ctcttccagccttccttcct-3'; *β-Actin rev*, 5'-agcactgtgttggcgtacag-3'. For serial transplantation experiments, Ct values for all

samples were normalized against β-Actin and DNA CNV was expressed relative to SW620 cells before xenograft implants (pre-xeno), which was set at 1.

For each qPCR assay, samples were analyzed in triplicate or quadruplicate.

**Statistics**

All *P* values were calculated based on a 2-tailed *t* test with GraphPad Prism 5 software. All sets of data met normal distribution and displayed homogenous variance. Values of $P<0.05$ were considered significant. In bar graphs, error bars represent standard error of the mean (SEM). The sample size of each experiment is specified in figure legends.

**RNA in situ hybridization (ISH)**

ISH was performed according to the Affymetrix QuantiGene ViewRNA ISH Cell Assay. Briefly, cells were fixed in 4% formaldehyde for 30 min at room temperature, permeabilized by pretreating in buffer solution for 5 min at room temperature and digested with protease for 10 min. Target probe sets were applied and hybridized to the cells by incubating for 3 h at 40°C. Probes were used at a dilution of 1:50 for Type 1 HSATII (Affymetrix) and 1:100 for Type 6 GAPDH (Affymetrix). Signal was amplified through the sequential hybridization of PreAmplifier and Amplifer mixes to the target probe set and target RNA molecules were detected by applying Label Probe Mix. Cells were then counterstained with DAPI (5 mg/ml, Life Technologies) and slides were mounted using ProLong Gold Antifade Reagent (Life Technologies). Fluorescence microscopy using a Nikon 90i was used to visualize target mRNAs. Type 1 probes were detected in the Cy3 (550 nm) channel and Type 6 probes in the Cy5 channel (650 nm). Merged images were generated using NIS-Elements software.

**Metaphase spreads and DNA FISH**

Cultured cells were trypsinized, resuspended in culture medium and kept in solution in a tube as control for xenograft-derived single cell suspensions. Both preparations were treated with Karyomax Colcemid (Gibco/Life Technologies) at a final concentration of 20 ng/ml. After 4 h of incubation at 37°C, cells were pelleted and resuspended in ice-cold hypotonic solution (0.56% KCl). After 6 min of incubation at room temperature and a 4 min spin at 1000 rpm, pellets were fixed with methanol:acetic acid (3:1), incubated for 10 min at room temperature, spun again and resuspended in 1 ml fixative. Cells were finally dropped on wet slides from a height of about 45 cm. After "aging" the chromosomes for 72 hours, slides were washed once with 2X SSC for 10 minutes, dehydrated in 70, 90, 100% ethanol for 2 min each and air dried. Probe (Sat.2 Repeat LNA fluorescein probe, Exiqon) was added at a concentration of 1.25 mM in a 50% formamide/2X SSC pH 7.0/10% dextran sulphate solution. DNA was denatured at 75°C for 5 minutes, then hybridized for 4 h at 37°C. After three washes in 0.1X SSC at 50°C for 5 min, two washes in 4X SSC/0.05% Tween at 37°C for 5 min and one wash in PBS at room temperature for 5 min, slides were mounted with Vectashield Mounting Medium with DAPI (Vectorlabs). Pictures were taken at 100X magnification with a Nikon 90i scope with colour camera.

**Single molecule sequencing and data analysis**

Purified RNA was subjected to digital gene expression (DGE) sample prepping and analysis on the HeliScope Single Molecule Sequencer from formerly Helicos BioSciences and now SeqLL. This method has been previously described[28]. Briefly, single stranded cDNA was reverse transcribed from RNA with a dTU25V primer and the Superscript III cDNA synthesis kit (Invitrogen/Life Technologies). Purified single stranded cDNA was denatured and then a poly-A tail was added to the 3' end using terminal transferase (New England Biolabs).

Purified DNA was subjected to DNA sequencing sample prepping protocol from Helicos as previously described[29]. Briefly, genomic DNA was sheared with a Covaris S2 acoustic sonicator producing fragments averaging 200 bps and ranging from 100 to 500 bps. Cleaned-up DNA was then denatured and a poly-A tail was added to the 3' end using terminal transferase (New England Biolabs).

Tailed cDNA or DNA were then hybridized to the sequencing flow cell followed by "Fill and Lock" and single molecule sequencing. Sequence reads were subjected to filtering for a minimum read length of 25 and removal of artifact reads followed by alignment of reads to the human genome (UCSC hg19) as well as the satellite reference library from Repeat Masker[5] using the indexDPgenomic aligner[28]. RNA and genomic DNA sequence reads aligning to satellites were then normalized to total genomic alignments to obtain reads per million (rpm).

**Estimating satellite copy-number changes from whole genome sequencing**

We downloaded a total of 38 pairs (tumour and matched tissue or blood samples) of colorectal whole genome sequencing (WGS) data from The Cancer Genomics Hub (CGHub; https://cghub.ucsc.edu/), and dbGAP (accession ID: phs000374.v1.p1). Among these, 29 pairs were sequenced by The Cancer Genome Atlas (TCGA) project, and 9 pairs were generated by a previous study[21]. To detect amplification or deletion of satellite DNA in the high-coverage (> 30X) cancer WGS data, we developed a computational method, called *repeat-CNV* (repeat copy number variation; *manuscript in preparation*). Briefly, we built a repeat sequence library that consists of consensus sequences of satellite repeats from Repbase (http://www.girinst.org/repbase/) and satellite repeat sequences annotated in the human genome reference by RepeatMasker (http://www.repeatmasker.org/)[30]. All of the sequencing reads were mapped to the

custom-built repeat sequence library using a revised *bwa* algorithm to assure that each read was uniquely associated with one satellite repeat[30]. The degree of enrichment and depletion of the satellite repeats were estimated based on the ratio of total read counts uniquely mapped to the satellite repeat in a cancer genome versus its matched normal (tissue or blood) genome using the previously described statistics[30]. The read counts were adjusted for the GC-bias estimated from reads mapped to genomic copy-neutral regions and (optionally) for broad copy number change, such as aneuploidy (10 Mb copy number correction is shown in Extended Data Figure 5b). An improved version of BIC-seq[31], called *Integer* (*manuscript in review*) was used to estimate copy number as well as purity and ploidy of cancer genomes.

**REFERENCES**


1. Djebali, S.*, et al.* Landscape of transcription in human cells. *Nature* **489**, 101-108 (2012).
2. Ugarkovic, D. Functional elements residing within satellite DNAs. *EMBO Rep* **6**, 1035-1039 (2005).
3. Plohl, M., Luchetti, A., Mestrovic, N. & Mantovani, B. Satellite DNAs between selfishness and functionality: structure, genomics and evolution of tandem repeats in centromeric (hetero)chromatin. *Gene* **409**, 72-82 (2008).
4. Bierhoff, H., Postepska-Igielska, A. & Grummt, I. Noisy silence: Non-coding RNA and heterochromatin formation at repetitive elements. *Epigenetics* **9**, 53-61 (2013).
5. Ting, D.T.*, et al.* Aberrant overexpression of satellite repeats in pancreatic and other epithelial cancers. *Science* **331**, 593-596 (2011).
6. Ozsolak, F.*, et al.* Amplification-free digital gene expression profiling from minute cell quantities. *Nat Methods* **7**, 619-621 (2010).
7. Jeanpierre, M. Human satellites 2 and 3. *Ann Genet* **37**, 163-171 (1994).
8. Bouzinba-Segard, H., Guais, A. & Francastel, C. Accumulation of small murine minor satellite transcripts leads to impaired centromeric architecture and function. *Proc Natl Acad Sci U S A* **103**, 8709-8714 (2006).
9. Rizzi, N.*, et al.* Transcriptional activation of a constitutive heterochromatic domain of the human genome in response to heat shock. *Mol Biol Cell* **15**, 543-551 (2004).
10. Azzalin, C.M., Reichenbach, P., Khoriauli, L., Giulotto, E. & Lingner, J. Telomeric repeat containing RNA and RNA surveillance factors at mammalian chromosome ends. *Science* **318**, 798-801 (2007).

11. Lander, E.S.*, et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).

12. Kazazian, H.H., Jr. & Goodier, J.L. LINE drive. retrotransposition and genome instability. *Cell* **110**, 277-280 (2002).

13. Boguslawski, S.J.*, et al.* Characterization of monoclonal antibody to DNA.RNA and its application to immunodetection of hybrids. *J Immunol Methods* **89**, 123-130 (1986).

14. Huertas, P. & Aguilera, A. Cotranscriptionally formed DNA:RNA hybrids mediate transcription elongation impairment and transcription-associated recombination. *Mol Cell* **12**, 711-721 (2003).

15. Hu, Z., Zhang, A., Storz, G., Gottesman, S. & Leppla, S.H. An antibody-based microarray assay for small RNA detection. *Nucleic Acids Res* **34**, e52 (2006).

16. Rigby, R.E.*, et al.* RNA:DNA hybrids are a novel molecular pattern sensed by TLR9. *EMBO J* **33**, 542-558 (2014).

17. Kukhanova, M.*, et al.* L- and D-enantiomers of 2',3'-dideoxycytidine 5'-triphosphate analogs as substrates for human DNA polymerases. Implications for the mechanism of toxicity. *J Biol Chem* **270**, 23055-23059 (1995).

18. Dai, L., Huang, Q. & Boeke, J.D. Effect of reverse transcriptase inhibitors on LINE-1 and Ty1 reverse transcriptase activities and on LINE-1 retrotransposition. *BMC Biochem* **12**, 18 (2011).

19. Lee, E.*, et al.* Landscape of somatic retrotransposition in human cancers. *Science* **337**, 967-971 (2012).

20. Altemose, N., Miga, K.H., Maggioni, M. & Willard, H.F. Genomic characterization of large heterochromatic gaps in the human genome assembly. *PLoS Comput Biol* **10**, e1003628 (2014).

21. Bass, A.J.*, et al.* Genomic sequencing of colorectal adenocarcinomas identifies a recurrent VTI1A-TCF7L2 fusion. *Nat Genet* **43**, 964-968 (2011).

22. Jiang, J., Birchler, J.A., Parrott, W.A. & Dawe, R.K. A molecular view of plant centromeres. *Trends Plant Sci* **8**, 570-575 (2003).

23. Neumann, P.*, et al.* Plant centromeric retrotransposons: a structural and cytogenetic perspective. *Mob DNA* **2**, 4 (2011).

24. Carone, D.M.*, et al.* A new class of retroviral and satellite encoded small RNAs emanates from mammalian centromeres. *Chromosoma* **118**, 113-125 (2009).

25. Ganem, N.J. & Pellman, D. Linking abnormal mitosis to the acquisition of DNA damage. *J Cell Biol* **199**, 871-881 (2012).

26. Crosetto, N.*, et al.* Nucleotide-resolution DNA double-strand break mapping by next-generation sequencing. *Nat Methods* **10**, 361-365 (2013).

27. Martinez, A.C. & van Wely, K.H. Centromere fission, not telomere erosion, triggers chromosomal instability in human carcinomas. *Carcinogenesis* **32**, 796-803 (2011).

28. Lipson, D.*, et al.* Quantification of the yeast transcriptome by single-molecule sequencing. *Nat Biotechnol* **27**, 652-658 (2009).

29. Pushkarev, D., Neff, N.F. & Quake, S.R. Single-molecule sequencing of an individual human genome. *Nat Biotechnol* **27**, 847-850 (2009).

30. Day, D.S., Luquette, L.J., Park, P.J. & Kharchenko, P.V. Estimating enrichment of repetitive elements from high-throughput sequence data. *Genome Biol* **11**, R69 (2010).

31. Xi, R.*, et al.* Copy number variation detection in whole-genome sequencing data using the Bayesian information criterion. *Proc Natl Acad Sci U S A* **108**, E1128-1136 (2011).

**FIGURES**

**Figure 1** | **HSATII is expressed in human tumours and 3D cancer cell models. a**, An outline of the experimental layout. **b,** Northern blot analysis of HSATII expression in tissue samples (Normal, normal pancreas; PDAC, pancreatic ductal adenocarcinoma; Duodenal ca., duodenal carcinoma; Well diff NET, well differentiated neuroendocrine tumour). Reads per million (rpm) assessed by DGE profiling are indicated below. **c**, Linear regression of DGE and Northern blot data (quantified by densitometry) of the same samples. **d**, Northern blot analysis of HSATII expression in HCT116 cells grown as 2D cultures or xenografts. **e**, Northern blot analysis of HSATII expression in colon cancer cell lines grown as 2D cultures or tumour spheres (3D). Ethidium bromide (Et Br) stainings of gels or U6-probed blots are shown for each Northern analysis as loading controls. **f**, RNA in situ hybridization (with the indicated fluorescent probes) of SW620 cells cultured under 2D conditions or as tumour spheres (60X magnification).

**Figure 2** | **HSATII RNA gives rise to DNA:RNA hybrids whose production is affected by reverse transcription inhibition. a,** Northern blot analysis of HSATII in Untreated (NT) or DNase I-treated extracts obtained from SW620 xenografts. Numbers below indicate relative signal quantitation. **b,** Extracts obtained from cells transfected with HSATII/GFP IVT products were subjected to RNase H treatment followed by Northern blot and hybridization to detect the RNA (HSATII S) strand of the hybrid. **c,** Northern blot analysis of extracts from 293T cells either untransfected or transfected with IVT HSATII or GFP, subjected to indicated nuclease treatment and probed for HSATII S. **d,** Northern blot analysis of extracts from 293T cells after transfection with IVT GFP, treated with RNase H and probed for GFP S. Ethidium bromide (Et Br) stainings of gels are shown as loading controls. **e,** Total nucleic acids (TNA) were isolated from IVT

HSATII-transfected 293T cells or COLO205 cells cultured in the presence of ddC or DMSO, followed by complete DNAse I digestion. After RNase H treatment, DNA:RNA hybrids were pulled down using a hybrid-specific antibody and their relative quantities were measures by HSATII-chr10 real-time PCR (qPCR). **f,** Fold change in the enrichment of DNA:RNA hybrids in HSATII-transfected 293T cells measured by qPCR after DRIP. **g,** Fold enrichment of endogenous HSATII DNA:RNA hybrids in COLO205 cells analysed by HSATII-chr10 real-time PCR after DRIP. For all charts, values represent the average of two independent experiments performed in triplicate ±SEM. **P*<0.05 (*t* test).

**Figure 3** | **HSATII rdDNA is re-integrated at the same original loci in the genome leading to pericentromere elongation in colon cancer xenografts. a,** DGE (RNA) and copy number (gDNA) analysis of satellite repeats (HSATII, GSATII) in the indicated samples (SW620) quantitated by single molecule sequencing. **b,** Representative HSATII DNA FISH on metaphase spreads of pre-xenograft 2D cultures and xenografts obtained from SW620 cells (100X magnification). **c,** Schematic representation of the genomic alignments of DNA-sequencing-derived reads, presented as fold change. **d, e, f** Copy number variation (CNV) in SW620 cells was assessed by real-time PCR on (**d**) HSATII-chr16-1 locus, (**e**) HSATII-chr16-2 locus and (**f**) chromosome 16q arm (T2, T6, T10 = 1 week of culture post-$2^{nd}$, -$6^{th}$ and -$10^{th}$ serial transplant, respectively). Error bars represent SEM of triplicate experiments. *P=t* test.

**Figure 4** | **Pericentromeric HSATII repeats expand both locally and genome-wide in primary human colon cancer samples. a, b,** Copy number variation (CNV) analysis of HSATII-chr16-1 (**a**) and HSATII-chr16-2 (**b**) loci on the indicated paired colon specimens (*n*=10). For each sample, values were normalized for β-Actin DNA and corrected for

chr16q arm changes. *P*=paired *t* test. **c,** Relative percentage of HSATII copy number changes in colon tumour/normal pairs according to combined HSATII-chr16-1 and -2 CNV analysis, including correction for chr16 arm gains/losses. **d, e,** Heat map of whole genome sequencing data (**d**) on the indicated primary colon cancer specimens and distribution based on a log2 ratio cut-off=0.1 (**e**).

**EXTENDED DATA**

**Extended Data Figure 1** | **HSATII expression in colon cancer cells is not induced by stress but by 3D culture conditions. a,** HCT116 cells were grown in low oxygen tension (1%) for the indicated time or UV irradiated for 15 minutes and let recover for 12 hours. RNA was extracted, run on a Northern blot and probed for HSATII. **b,** Analogous Northern blot on cells non treated (NT) or subjected to either heat shock (1 hour at 42°C followed by increasing recovery periods) or oxidative stress (200 mM $H_2O_2$ for the indicated times). **c,** Same Northern blot on either 2D HCT116 cell cultures or the same cells grown to confluence or in medium containing 5 μM 5-azacytidine for 72 h or in soft agar. **d,** Northern blot of HCT116 RNA extracted from the indicated culture conditions and probed for HSATII. For tumour spheres, co-culture with irradiated MEF feeder layer cells started 5 days after growth as spheres in absence of adhesion. **e,** Northern blot analysis of HSATII expression in SW620 cells before (2D), during (tum. spheres) and at the indicated times after (post-TS) culture under non-adherent conditions, then during (xeno) and at the indicated times after (post-xeno) *in vivo* growth. Post- time points reflect days after replating in 2D conditions. **f,** HSATII detection by Northern blot in SW620 cells, grown as 2D cultures or xenografts, on both sense (S) and antisense (AS) orientation. **g,** Northern blot following differential nuclear/cytoplasmic RNA extraction from HCT116 2D culture and xenograft shows nuclear localization of HSATII. U6-probed blots or ethidium bromide (Et Br) stainings are shown for each Northern as loading controls.

**Extended Data Figure 2** | **HSATII signal is not deriving from a genomic DNA component. a,** Trizol extracts of the same SW620 tumour xenografts, processed in parallel using either regular (REG) or Phase Lock Gel (PLG) tubes, were run on a gel,

blotted and probed for HSATII. **b,** RNA and gDNA extracts from SW620 xenografts were run on a Northern blot. Ethidium bromide (Et Br) staining of gels was used as loading control.

**Extended Data Figure 3 | IVT reaction specifically generates single stranded HSATII RNA. a,** Agarose gel electrophoresis of HSATII-chr10 fragment before (PCR) and after IVT with (+T7) or without (-T7) addition of RNA polymerase. **b,** *In vitro* transcribed HSATII RNA was run on a Northern blot and probed for HSATII on both sense (S) and antisense (AS) strand. **c,** RNA in situ hybridization (with the indicated fluorescent probes) of 293T cells following transfection of IVT HSATII (20X magnification). **d,** *In vitro* transcribed HSATII or GFP were transfected into 293T cells, and extracted with Trizol 24 h later. Extracts were treated with either DNase I or left untreated (NT) and probed for HSATII S to detect HSATII RNA. **e,** Northern blot of extracts from 293T cells transfected with IVT HSATII or GFP, subjected to the indicated nuclease treatment and probed for HSATII S. No T7 polymerase (-T7) was added to the control IVT reaction. **f,** Northern blot of extracts from 293T cells transfected with IVT GFP, treated as indicated in the figure and probed with a GFP S-specific oligonucleotide. Ethidium bromide (Et Br) stainings of gels are shown as loading controls. **g,** HSATII PCR product was transfected in 293T cells, followed by nucleic acid extraction 24 h later. Extracts were either treated with DNase I or left untreated (NT) and probed for HSATII S and AS. **h,** Northern blot shows no signal on either strand upon DNase I digestion of the extracts obtained as described above. **i,** Real-time PCR analysis on HSATII-chr10 locus following RNase A treatment of extracts obtained from COLO205 cells treated with DMSO or ddC for two weeks. Values represent average of three independent experiments ±SEM. *P*=*t* test.

**Extended Data Figure 4 | HSATII DNA amplification is acquired in cancer cell xenografts.** Schematic representation of the genomic alignments of DNA-sequencing-derived reads, presented as absolute reads per million (rpm), in the two indicated experimental conditions (SW620 cells).

**Extended Data Figure 5 | Pericentromeric HSATII repeats are expanded in primary human kidney cancer samples and satellite DNA enrichment in colorectal cancer samples was adjusted for large copy number variations (CNV). a,** Pie chart displaying percentages of HSATII copy number changes in kidney tumour/normal pairs according to combined HSATII-chr16-1 and -2 copy number variation (CNV) analysis, including correction for chr16 arm gains/losses. **b,** Only the 31 samples for which genomic copy number profiles could be confidently estimated were included in the analysis. Specifically, the CNV correction analysis was performed for 14 out of 20 samples with HSATII gain (before CNV correction), 8 out of 8 samples with no significant HSATII change, and 9 out of 10 with HSATII loss. The genomes of the other samples showed abundant variations, making it difficult to distinguish copy-neutral and altered regions. After correcting the satellite DNA enrichment by that expected from the detected CNV (10 Mb and larger), all the samples (•) except two (✖) showed the same amplification/depletion trend as reported in Figure 4d. After the correction, the two exceptions went from HSATII gain to no change (tcga_3896) and from no change to HSATII loss (tcga_A01T).

**Extended Data Figure 6 | Proposed model of HSATII overexpression leading to progressive amplification of pericentromeric loci through a reverse transcription mechanism in 3D cancer cell cultures, xenografts and human tumours.** Anchorage-independent culture conditions as well as *in vivo* growth lead to derepression of HSATII

pericentromeric repeats in human cancer cells. Once expressed, these sequences undergo reverse transcription that mediates the formation of DNA:RNA hybrids, which eventually give rise to RNA-derived DNA (rdDNA) intermediates. In turn, these molecules display the potential to be reintegrated into the genome within the same pericentromeric regions from which they originally derive. This process gradually induces an expansion of pericentromeric loci in cancer, with important implications for the maintenance of genomic integrity and chromosomal stability in tumour cells.

**Extended Data Table 1** | **Comprehensive summary of all satellite repeats identified by DNA- and RNA-seq in the indicated samples derived from SW620 cells.** Values are expressed as reads per million (rpm).

Figure 1

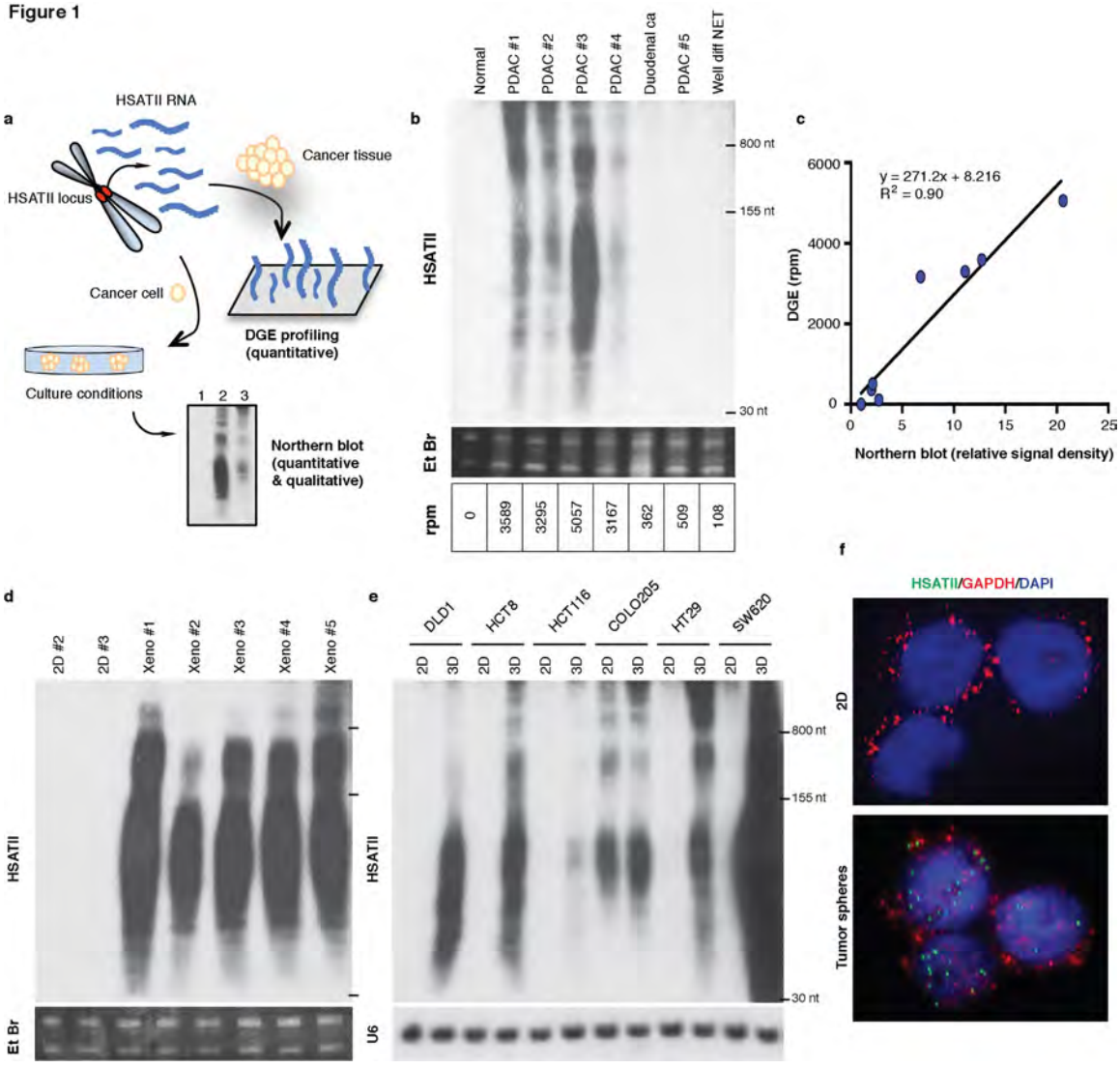

a
HSATII RNA
HSATII locus
Cancer tissue
Cancer cell
DGE profiling
(quantitative)
Culture conditions
1 2 3
Northern blot
(quantitative
& qualitative)
b
Normal
PDAC #1
PDAC #2
PDAC #3
PDAC #4
Duodenal ca
PDAC #5
Well diff NET
HSATII
800 nt
155 nt
30 nt
Et Br
rpm
0
3589
3295
5057
3167
362
509
108
c
y = 271.2x + 8.216
R2 = 0.90
DGE (rpm)
Northern blot (relative signal density)
d
2D #2
2D #3
Xeno #1
Xeno #2
Xeno #3
Xeno #4
Xeno #5
HSATII
Et Br
e
DLD1
HCT8
HCT116
COLO205
HT29
SW620
2D
3D
HSATII
800 nt
155 nt
30 nt
U6
f
HSATII/GAPDH/DAPI
2D
Tumor spheres

Figure 2

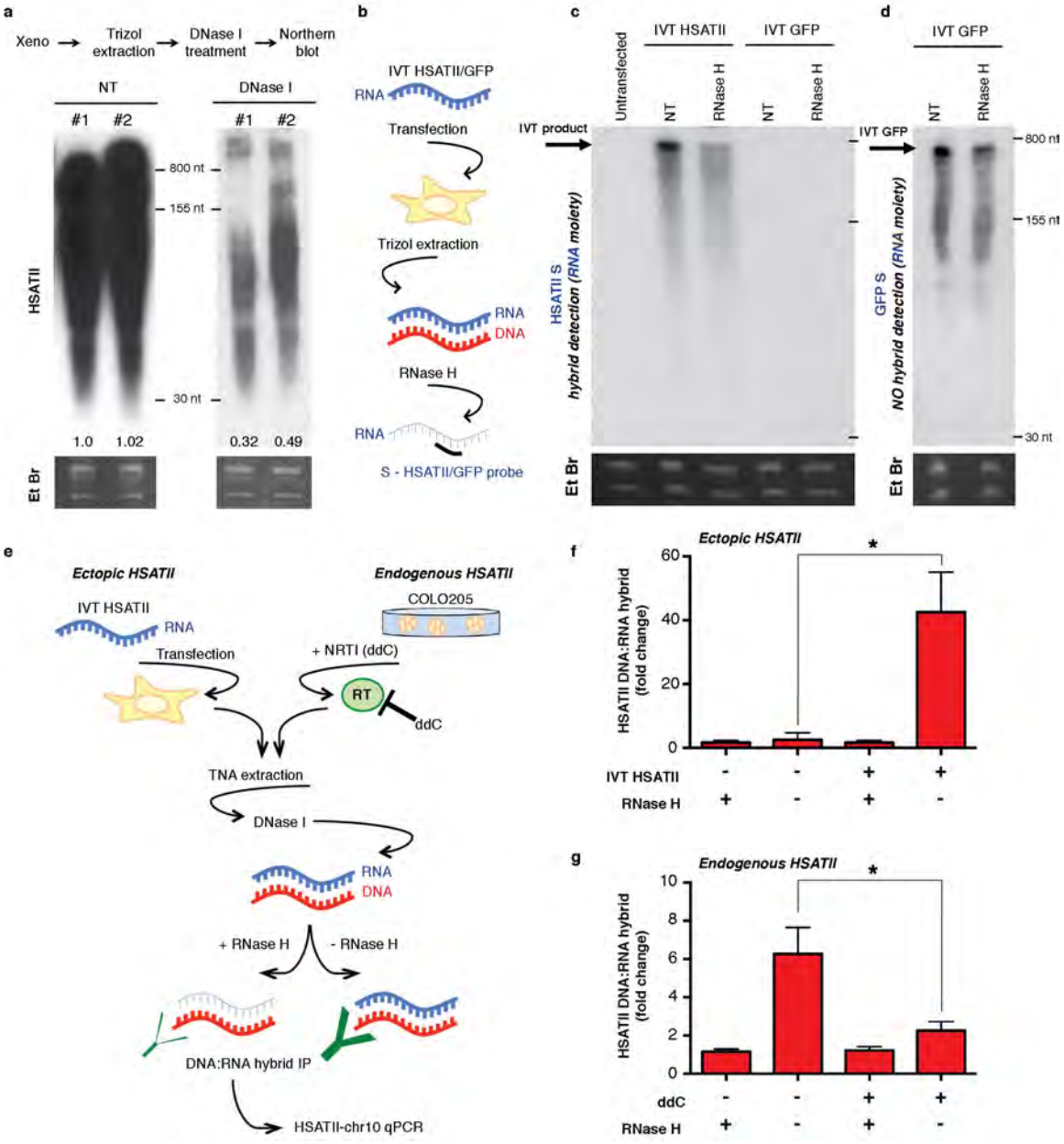

a
Xeno
Trizol extraction
DNase I treatment
Northern blot
NT
DNase I
#1 #2 #1 #2
800 nt
155 nt
30 nt
HSATII
1.0 1.02 0.32 0.49
Et Br
b
IVT HSATII/GFP
RNA
Transfection
Trizol extraction
RNA
DNA
RNase H
RNA
S - HSATII/GFP probe
c
IVT HSATII
IVT GFP
Untransfected
NT
RNase H
NT
RNase H
IVT product
HSATII S
hybrid detection (RNA moiety)
Et Br
d
IVT GFP
NT
RNase H
IVT GFP
800 nt
155 nt
30 nt
GFP S
NO hybrid detection (RNA moiety)
Et Br
e
Ectopic HSATII
Endogenous HSATII
IVT HSATII
RNA
COLO205
Transfection
+ NRTI (ddC)
RT
ddC
TNA extraction
DNase I
RNA
DNA
+ RNase H
- RNase H
DNA:RNA hybrid IP
HSATII-chr10 qPCR
f
Ectopic HSATII
*
HSATII DNA:RNA hybrid (fold change)
60
40
20
0
IVT HSATII - - + +
RNase H + - + -
g
Endogenous HSATII
*
HSATII DNA:RNA hybrid (fold change)
10
8
6
4
2
0
ddC - - + +
RNase H + - + -

Figure 3

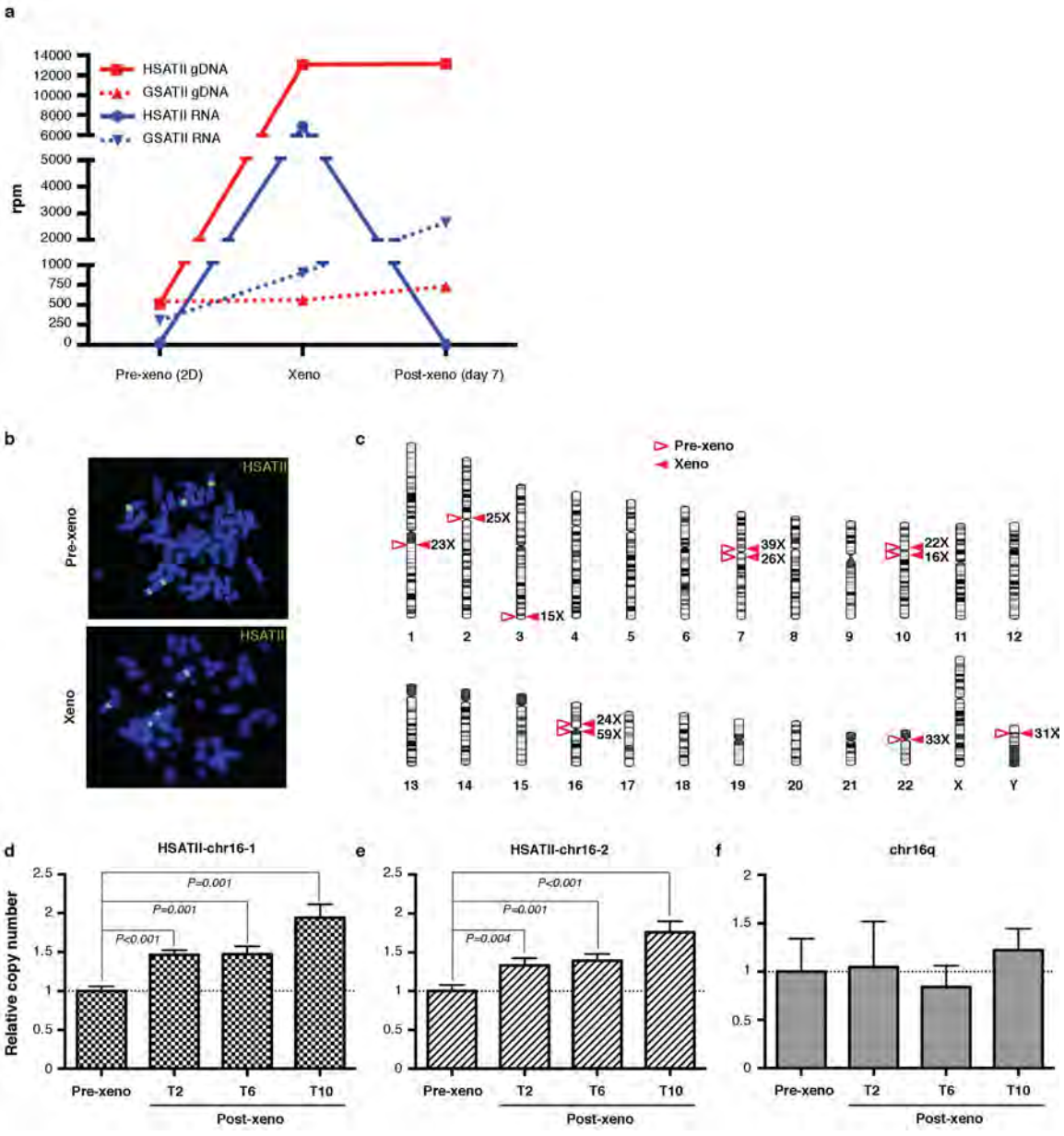
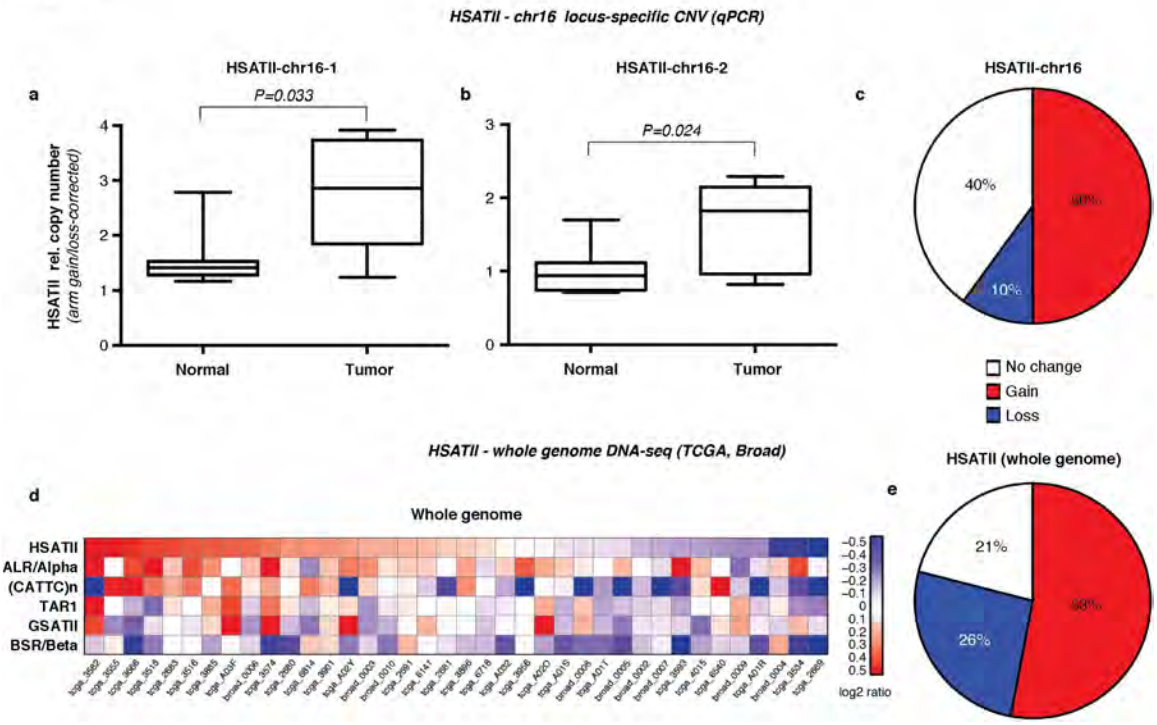
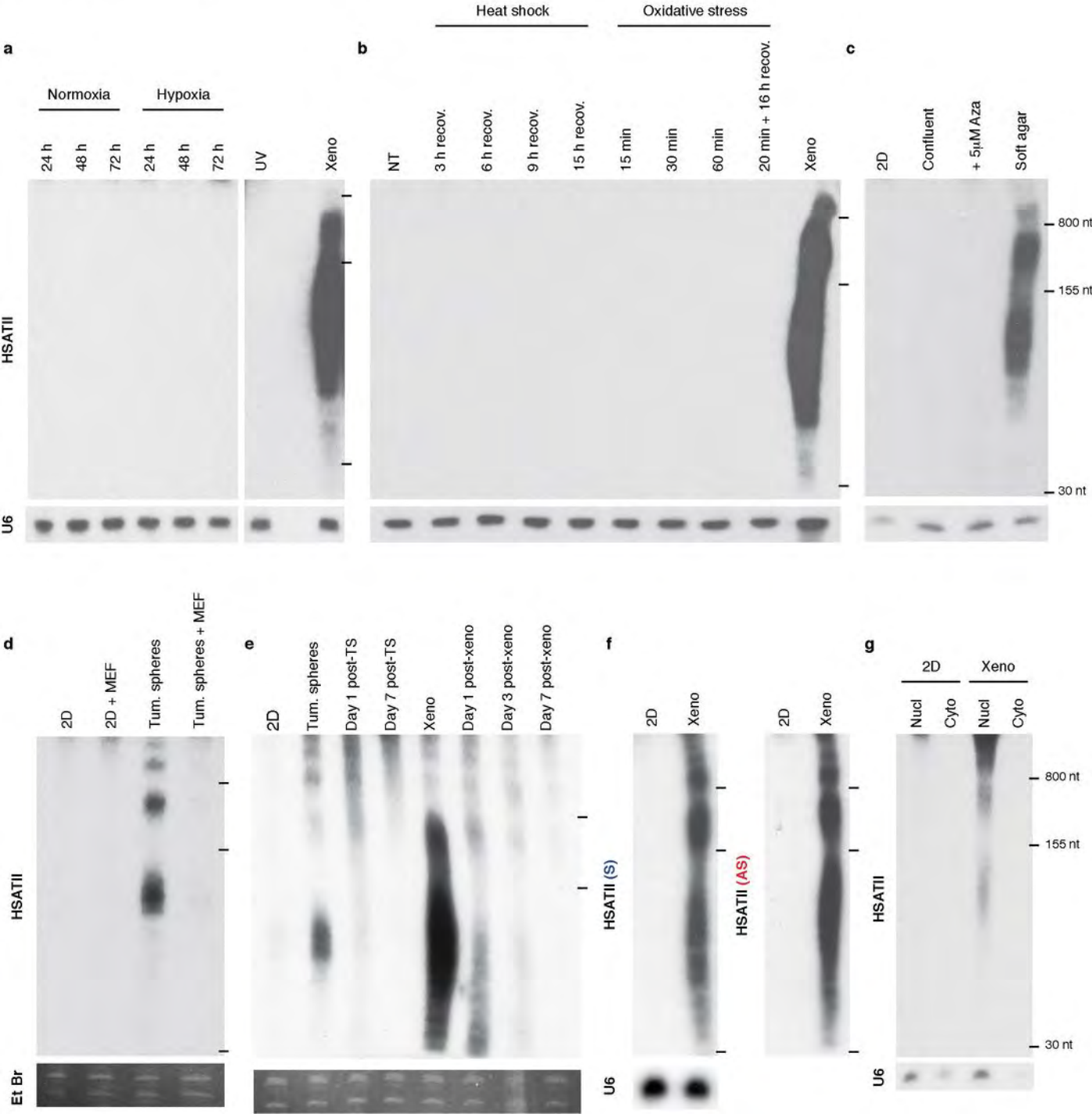

a
HSATII gDNA
GSATII gDNA
HSATII RNA
GSATII RNA
rpm
Pre-xeno (2D)
Xeno
Post-xeno (day 7)
b
HSATII
Pre-xeno
Xeno
c
Pre-xeno
Xeno
23X
25X
15X
39X
26X
22X
16X
24X
59X
33X
31X
d
HSATII-chr16-1
Relative copy number
P=0.001
P=0.001
P<0.001
Pre-xeno
T2
T6
T10
Post-xeno
e
HSATII-chr16-2
P<0.001
P=0.001
P=0.004
f
chr16q

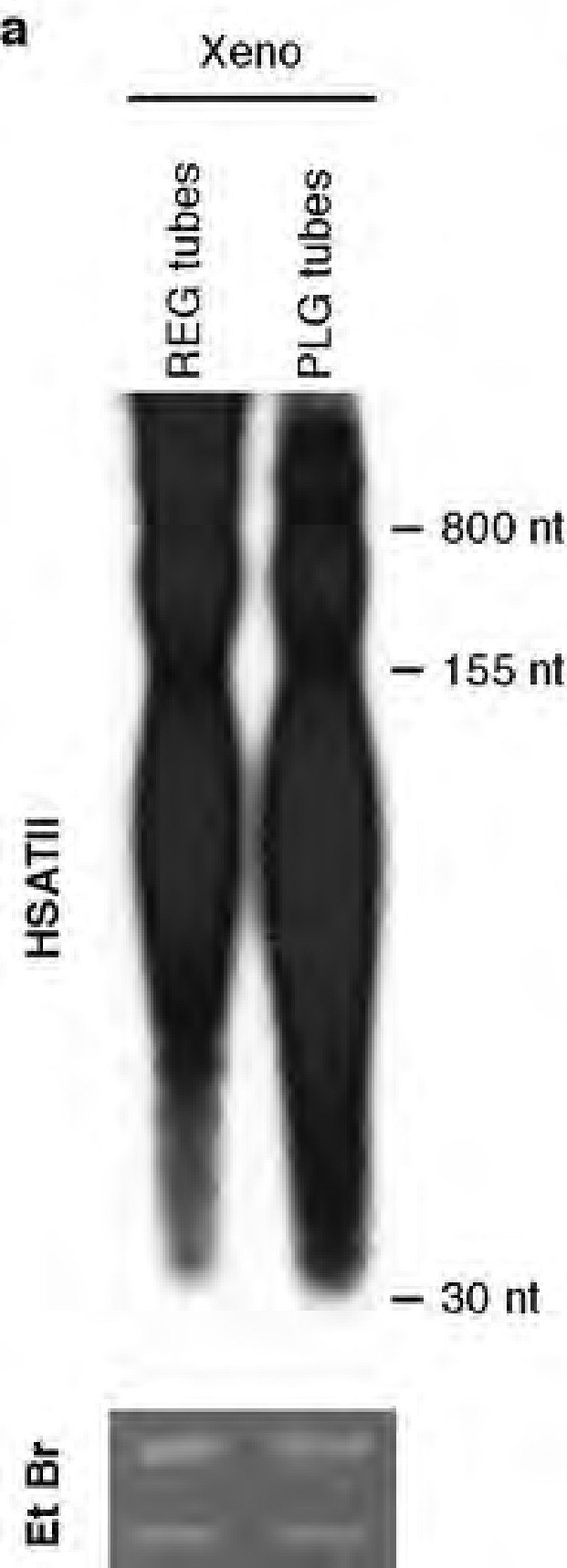
Figure 4

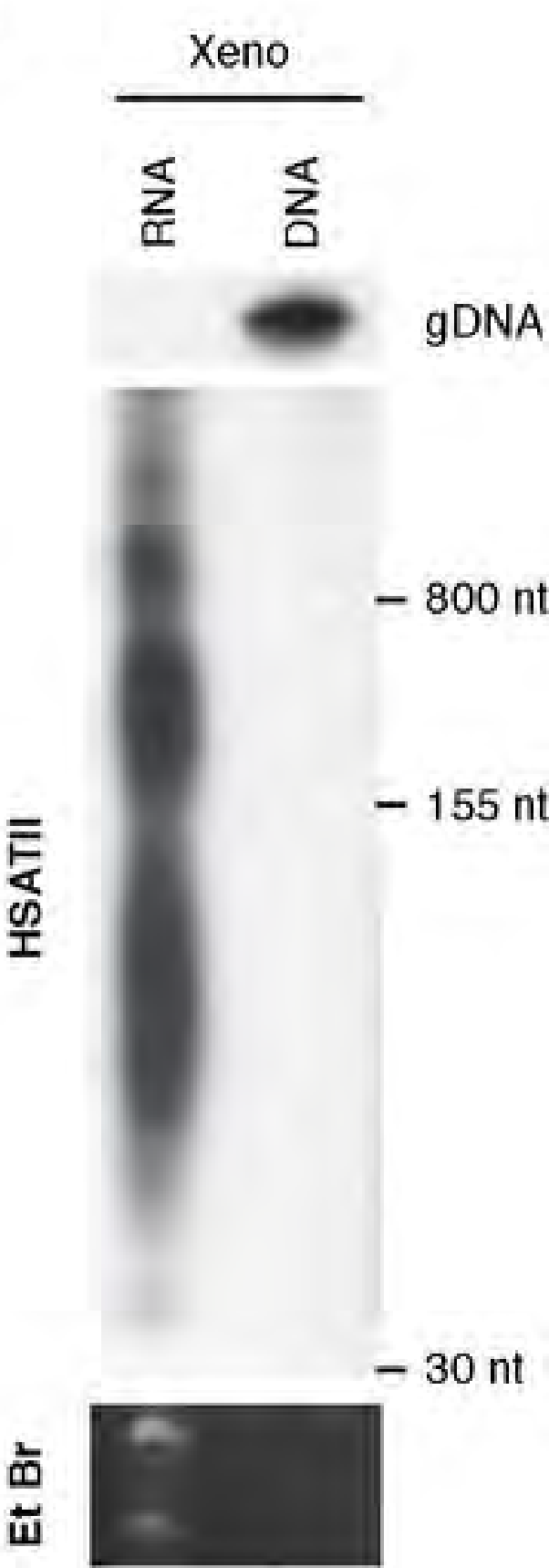

HSATII - chr16 locus-specific CNV (qPCR)
a
HSATII-chr16-1
P=0.033
HSATII rel. copy number
(arm gain/loss-corrected)
Normal
Tumor
b
HSATII-chr16-2
P=0.024
Normal
Tumor
c
HSATII-chr16
40%
10%
No change
Gain
Loss
HSATII - whole genome DNA-seq (TCGA, Broad)
d
Whole genome
HSATII
ALR/Alpha
(CATTC)n
TAR1
GSATII
BSR/Beta
log2 ratio
e
HSATII (whole genome)
21%
53%
26%

**Extended Data Figure 1**

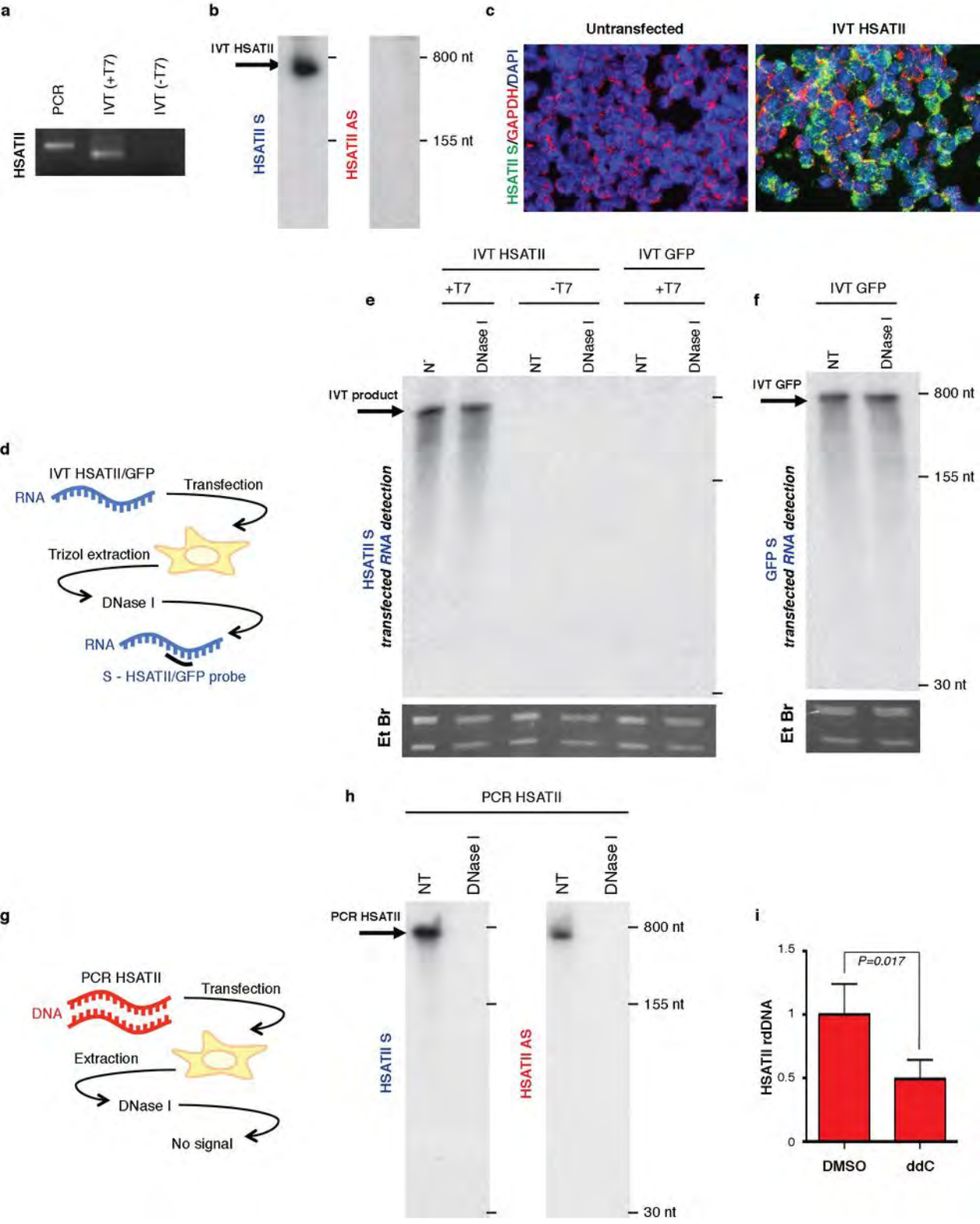


# Extended Data Figure 2

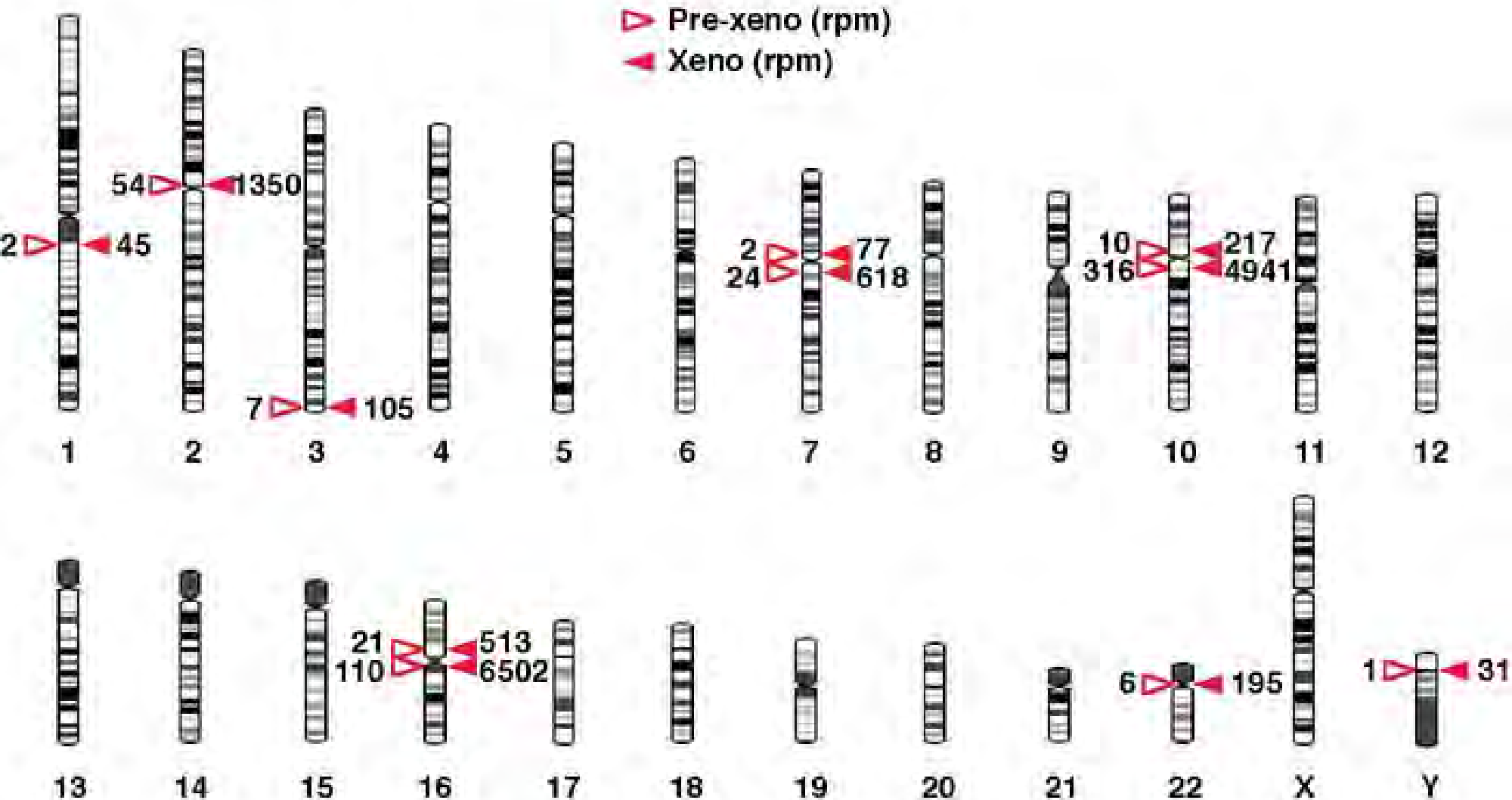


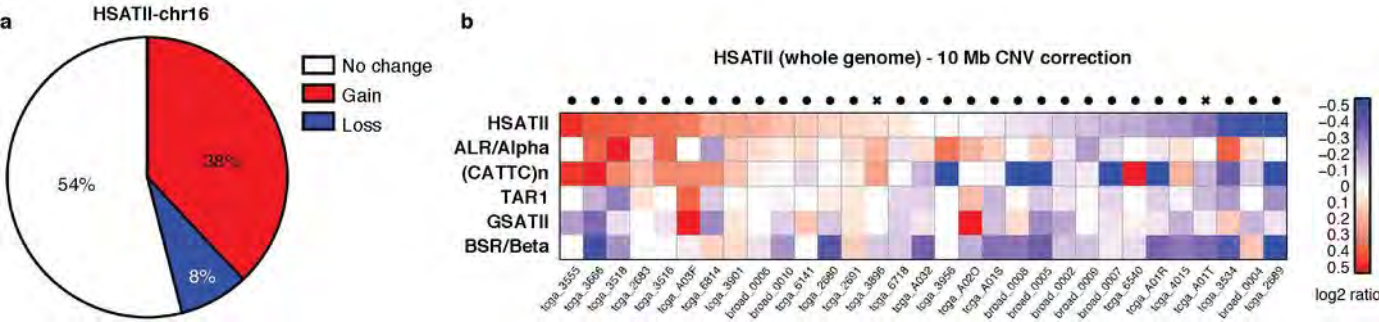
**Extended Data Figure 3**



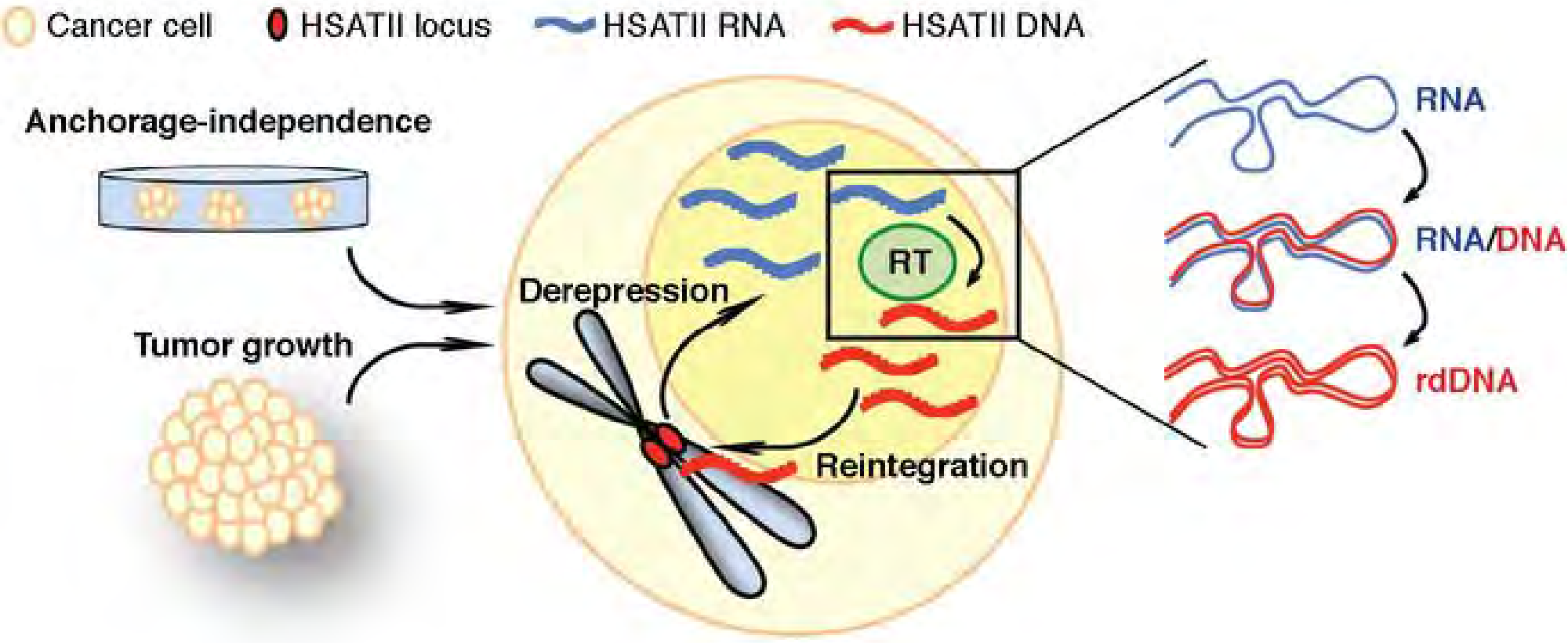
**Extended Data Figure 4**



**Extended Data Figure 5**



**Extended Data Figure 6**



**Extended Data Table 1**

| | *DNA CNV (rpm)* | | | *RNA DGE (rpm)* | | |
|---|---|---|---|---|---|---|
| Sat. type | Pre-xeno | Xeno | Post-xeno | Pre-xeno | Xeno | Post-xeno |
| ALR/Alpha | 4922 | 66549 | 72474 | 2371 | 52602 | 5932 |
| HSATII | 518 | 13066 | 13143 | 19 | 6908 | 4 |
| BSR/Beta | 855 | 4962 | 5077 | 712 | 3501 | 1883 |
| (GAATG)n | 113 | 3949 | 3379 | 25 | 2395 | 37 |
| (CATTC)n | 135 | 3604 | 3069 | 19 | 2044 | 54 |
| SAR | 226 | 1786 | 2211 | 0 | 782 | 4 |
| GSATII | 545 | 565 | 730 | 305 | 904 | 2662 |
| ACRO1 | 15 | 1183 | 707 | 203 | 849 | 158 |
| SST1 | 168 | 659 | 651 | 254 | 506 | 412 |
| SATR1 | 84 | 525 | 595 | 127 | 428 | 62 |
| CER | 66 | 632 | 588 | 64 | 412 | 37 |
| SATR2 | 43 | 361 | 374 | 133 | 267 | 58 |
| HSATI | 3 | 230 | 216 | 0 | 192 | 4 |
| REP522 | 65 | 194 | 152 | 210 | 294 | 483 |
| GSAT | 44 | 112 | 112 | 127 | 140 | 233 |
| TAR1 | 643 | 113 | 104 | 235 | 209 | 187 |
| HSAT4 | 20 | 94 | 83 | 83 | 82 | 150 |
| D20S16 | 5 | 66 | 57 | 44 | 46 | 8 |
| LSAU | 16 | 59 | 53 | 19 | 46 | 37 |
| GSATX | 37 | 34 | 38 | 57 | 228 | 604 |
| MSR1 | 8 | 45 | 26 | 0 | 28 | 8 |
| HSAT5 | 2 | 37 | 25 | 13 | 32 | 4 |
| HSAT6 | 0 | 9 | 6 | 0 | 5 | 0 |
| SUBTEL_sa | 4 | 1 | 2 | 19 | 1 | 4 |

## SUPPLEMENTARY INFORMATION

### Supplementary Data 1

First, to achieve a cleaner separation of the aqueous phase from the gDNA-containing interphase, we used Phase Lock Gel (PLG) tubes, which separate the aqueous and organic phase with a solid barrier and prevent cross contamination. The PLG-based extraction conditions yielded the same result as standard Trizol extraction using regular (REG) tubes (Extended Data Fig. 2a). Second, we processed in parallel RNA and total gDNA from a colon cancer xenograft: HSATII gDNA was visible as a high molecular weight species trapped in the well of a 4% polyacrylamide-urea gel, whereas the medium/small molecular weight fraction nucleic acid was positive for HSATII signal on the Northern blot only following RNA extraction (Extended Data Fig. 2b). Third, we processed the same cell lines following culture under diverse conditions: only under 3D or xenograft conditions were medium/small molecular weight HSATII sequences evident by Northern blots (Fig. 1d, e and Extended Data Fig. 1). The rapid loss of HSATII signal following replating of 3D cultures into 2D, despite processing under identical experimental conditions, further excluded gDNA contamination as a source for the HSATII DNA signal on Northern blot (Extended Data Fig. 1e).

### Supplementary Data 2

First, no band was visualized in the HSATII IVT product by agarose gel electrophoresis unless T7 polymerase was present in the transcription reaction (Extended Data Fig. 3a). Second, only the sense strand (S), corresponding to the transcribed HSATII RNA was evident by Northern blot of the IVT product, and no signal was detectable on the antisense strand (AS; Extended Data Fig. 3b). Third, RNA ISH analysis of 293T cells (2D culture) transfected with ectopic single stranded IVT sense HSATII RNA showed

strong expression of the transcribed sense HSATII sequence with no nonspecific detection of gDNA signal (Extended Data Fig. 3c). Fourth, presence of the expected 670 nt sense (S) HSATII RNA was readily detected using a single stranded complementary probe in PLG Trizol extracts obtained from transfected cells with no loss of signal following DNase I digestion (Extended Data Fig. 3d, e). Furthermore, no HSATII band was detected using either sense or antisense probes in untransfected cells (Fig. 2c). In contrast, if the HSATII-chr10 PCR DNA template itself, rather than the purified IVT RNA, was directly transfected into 293T cells, signal by Northern blot was not strand-specific and was completely abrogated by DNase I treatment (Extended Data Fig. 3g, h). Finally, all experiments were performed with IVT Green Fluorescent Protein (GFP) as an additional control for RNA signal specificity (Extended Data Fig. 3e, f). Taken all together, our controls proved that single stranded sense HSATII RNA was successfully transfected into 293T cells, without contaminating template DNA.